

ZESZYTY NAUKOWE
WYDZIAŁU ETI POLITECHNIKI GDAŃSKIEJ

TECHNOLOGIE INFORMACYJNE



DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

NR 3



1904 1945 2004/2005
JUBILEUSZ POLITECHNIKI W GDAŃSKU

BEST AVAILABLE COPY
20041206 054

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 23 August 2004	3. REPORT TYPE AND DATES COVERED Conference Proceedings, 18 May 2004	
4. TITLE AND SUBTITLE 2 nd Conference on Information Technology / Special Session on Homeland Security, Volume 3			5. FUNDING NUMBERS FA8655-04-1-5047	
6. AUTHOR(S) Conference Committee				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Gdansk University of Technology ul. Gabriela Narutowicza 11/12 80-952 Gdansk 80-952 Poland			8. Performing Organization Report Number	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) EOARD PSC 802 Box 14 FPO 09499-0014			10. SPONSORING/MONITORING AGENCY REPORT NUMBER CSP 04-5047	
11. SUPPLEMENTARY NOTES Volume 3, ISBN 83-917681-5-5 Copyright 2004 Wydział ETI Politechniki Gdanskiej Gdansk. Available from: Wydział ETI Politechniki Gdanskiej, Gdansk. The Department of Defense has permission to use for government purposes only. All other rights are reserved by the copyright holder.				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited. (approval given by local Public Affairs Office)			12b. DISTRIBUTION CODE A	
ABSTRACT (Maximum 200 words) The Final Proceedings for 2 nd Conference on Information Technology / Special Session on Homeland Security, 16-18 May 2004 Formal methods in information engineering Electronic documents and digital libraries Autonomous robots in embeded systems Mobile and portable information systems Design and implementation of methodologies and technologies for information based products Dependability and security of information processing systems Development of IT infrastructure Special English Session on Homeland Security				
14. SUBJECT TERMS EOARD, Computational methods, C31, Computer network security				15. NUMBER OF PAGES 326
16. SECURITY CLASSIFICATION OF: a. Report UNCLASSIFIED b. Abstract UNCLASSIFIED c. This page UNCLASSIFIED		17. LIMITATIONS OF ABSTRACT UNCLASSIFIED	18a. NAME OF RESPONSIBLE PERSON Paul Losiewicz, Ph. D. 18b. TELEPHONE NUMBER (include area code) +44 20 7514 4474	

**ZESZYTY NAUKOWE
WYDZIAŁU ETI POLITECHNIKI GDAŃSKIEJ**

TECHNOLOGIE INFORMACYJNE

NR 3

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited



Gdańsk 2004

AQ F05-02-0395

KOMITET REDAKCYJNY

**Zeszytów Naukowych Wydziału ETI
Politechniki Gdańskiej**

Redaktorzy wydania

Józef Woźniak
Krzysztof Nowicki

Wydano za zgodą
DZIEKANA WYDZIAŁU ETI PG

W materiałach konferencyjnych zamieszczono
wyłącznie artykuły recenzowane

© Copyright by Wydział ETI Politechniki Gdańskiej
Gdańsk 2004

ISBN 83-917681-5-5

Druk: Zakład Poligrafii Politechniki Gdańskiej
ul. G. Narutowicza 11/12, 80-952 Gdańsk, tel. (0-58) 347 25 35

Technologie informacyjne (TI) umożliwiają implementację różnego typu systemów, szeroko stosowanych we współczesnym świecie. Nie oznacza to wcale, że są to już technologie dojrzałe, a ich wprowadzenie nie wiąże się z żadnym większym ryzykiem. Ogromne tempo rozwoju TI wynika z faktu, że korzyści z ich wykorzystania bądź nadzieja na osiągnięcie znacznych korzyści w przyszłości są ciągle duże. Na ogół konwergencja (integracja wielu istniejących rozwiązań) i synergia (generowanie nowych możliwości po integracji tych rozwiązań) są motorem ciągłego postępu.

II Konferencja Technologie Informacyjne jest zorganizowana przez Wydział Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej pod patronatem trzech Prezydentów Trójmiasta: Gdańska, Gdyni, Sopotu i J. M. Rektora PG w Roku Jubileuszowym Politechniki Gdańskiej, w rocznicę 100-lecia politechniki w Gdańsku. Odbywa się także już po fakcie ponownego włączenia się Polski do rodziny Krajów Europejskich. Stąd też większy zakres tematyczny tej konferencji, jak również częściowo międzynarodowy charakter. Znacznie powiększył się także skład Komitetu Programowego oraz liczba przyjętych do druku artykułów.

Tematy konferencji skupią się wokół 13 sesji tematycznych, w tym kilka sesji jest zorganizowanych przez uczestników konferencji. Są to: Homeland Security, Nietechniczne aspekty TI, Systemy mikroelektroniczne, Systemy radiowe. Tego typu sesje specjalistyczne chcemy rozwijać w przyszłości.

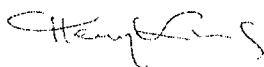
W pierwszym dniu Konferencji odbędzie się również specjalna sesja z okazji Światowego Dnia Telekomunikacji poświęcona problemom kształcenia, a wieczorem na bankiecie zostanie wręczony „Laur dla Pracodawcy” najlepszej firmie wybranej przez studentów, pracowników i absolwentów Wydziału ETI. W drugim dniu konferencji zorganizowane będą studenckie Targi Pracy. Odbędą się również dwie sesje plenarne, na których wybitni specjaliści przedstawiają bardzo interesujące referaty:

1. Prof. Wojciech Szpankowski: *Ubiquitous Pattern Matching and its Applications*,
2. Prof. Józef Lubacz i prof. Andrzej Krasniewski: *Ewolucja szkolnictwa wyższego w Europie i jej konsekwencje dla kształcenia w obszarze technik i technologii informacyjnych*.

Żałuję, że brakuje sesji poświęconej Twórcy Szkoły Mikrofal na Wydziale ETI PG Profesorowi Krzysztofowi Grabowskiemu. Niech więc wysoki poziom obrad oraz wyniki konferencji będą wyrazem podziękowania za Jego trud poświęcony rozwojowi naszego Wydziału.

Mam nadzieję, że obrady II Konferencji Technologie Informacyjne przyczynią się do lepszego wzajemnego poznania się wielu ekspertów z tej dziedziny. Wówczas dzięki skutkom konwergencji i synergii osiągniemy znacznie większy postęp we współpracy na rzecz zaspokajania potrzeb społeczeństwa informacyjnego.

Z podziękowaniem za współpracę



Dziekan Wydziału ETI
Politechnika Gdańska

Komitet Naukowy

Przewodniczący

Prof. Henryk Krawczyk (WETI PG)

Wiceprzewodniczący

Prof. Józef Woźniak (WETI PG)

Sekretarz

Dr Krzysztof Nowicki (WETI PG)

Członkowie:

Prof. Daniel Bem (WE PWr)
Prof. Michał Białko (PAN)
Prof. Zdzisław Bubnicki (PAN)
Prof. Andrzej Czyżewski (WETI PG)
Prof. Władysław Findeisen (PAN)
Prof. Tadeusz Galanc (WliZ PWr)
Prof. Janusz Górski (WETI PG)
Prof. Tomasz Imieliński (RU USA)
Prof. Zygmunt Kitowski (AMW)
Prof. Bogdan Kosmowski (WETI PG)
Prof. Marek Kubale (WETI PG)
Prof. Krzysztof Kuchciński (LIT, SE)
Prof. Józef Lubacz (WETI PW)
Prof. Stanisław F. Łęgowski (UW, USA)
Prof. Jerzy Mazur (WETI PG)
Prof. Maciej Niedźwiecki (WETI PG)
Prof. Antoni Nowakowski (WETI PG)
Prof. Tadeusz Orzechowski (WEAliE AGH)
Prof. Zdzisław Pawlak (PAN)
Prof. Krzysztof Pawlikowski (CU, NZ)
Prof. Michał Polowczyk (WETI PG)
Prof. Andrzej Ruciński (UNH, USA)
Prof. Dominik Rutkowski (WETI PG)
Prof. Jerzy Rutkowski (WAEil PŚI.)
Prof. Roman Salamon (WETI PG)
Prof. Wojciech Sobczak (WETI PG)
Prof. Andrzej Stepnowski (WETI PG)
Prof. Wojciech Szpankowski (PU, USA)
Prof. Jan Węglarz (PAN)
Prof. Stefan Węgrzyn (PAN)
Prof. Bogdan M. Wilamowski (AU, USA)
Prof. Bogdan Wiszniewski (WETI PG)
Prof. Wiesław Woliński (PAN)
Prof. Romuald Zielonko (WETI PG)
Prof. Marian Zientalski (WETI PG)
Prof. Jacek M. Żurada (UL, USA)

Recenzenci

Andrzej Czyżewski
Krzysztof Goczyła
Ewa Hermanowicz
Wojciech Jędruch
Jerzy Kaczmarek
Sylwester Kaczmarek
Renata Kalicka
Ryszard Katulski
Bogdan Kosmowski
Bożena Kostek
Henryk Krawczyk
Marek Kubale
Jerzy Mazur
Maciej Niedźwiecki
Antoni Nowakowski
Michał Polowczyk
Andrzej Ruciński
Dominik Rutkowski
Roman Rykaczewski
Wojciech Sobczak
Bogdan Wiszniewski
Józef Woźniak
Romuald Zielonko
Marian Zientalski

SPIS TREŚCI

EWOLUCJA SZKOLNICTWA WYŻSZEGO W EUROPIE I JEJ KONSEKWENCJE DLA KSZTAŁCENIA W OBSZARZE TECHNIK I TECHNOLOGII INFORMACYJNYCH Andrzej Kraśniewski, Józef Lubacz	1
UBIQUITOUS PATTERN MATCHING AND ITS APPLICATIONS Wojciech Szpankowski	17
SYNERGIA I KONERGENCJA PODSTAWĄ ROZWOJU NOWOCZESNYCH TECHNOLOGII INFORMACYJNYCH Henryk Krawczyk, Józef Woźniak	19
FRAUDULENT CONSUMER BEHAVIOR ANALYSIS AND DETECTION FOR UTILITY COMPANIES Péter Arató, Bálint Kiss, László Vajta, and Gábor Vámos	37
THE CONCEPT OF SMART AND SECURE LABORATORY Piotr Brudło	45
BEZPIECZNE SIECI VLAN Z AUTORYZACJĄ DOSTĘPU OPARTE NA CERTYFIKATACH ATRYBUTÓW Grzegorz Górski	51
PAKIET OCENY BEZPIECZEŃSTWA SYSTEMÓW INFORMACYJNYCH Henryk Krawczyk, Michał Wielgus	59
BIOWARFARE AGENT DETECTION WITH QUANTUM ENTANGLEMENT OF PHOTONS Henryk Malak	67
MECHANIZMY BEZPIECZEŃSTWA W BEZPRZEWODOWYCH SIECIACH 802.11 Wojciech Neubauer, Józef Woźniak	73
KONCEPCJA ROZPROSZONEGO SYSTEMU ZABEZPIECZAJĄCEGO POJAZD SAMOCHODOWY OPARTEGO NA DEDYKOWANYCH MAGISTRALACH CYFROWYCH Marek Niedostatkiwicz	85
OPTOELEKTRONICZNE METODY OCHRONY INFRASTRUKTURY TELEINFORMATYCZNEJ Jerzy Pluciński, Paweł Wierzbą	93
ZAGADNIENIA NIEZAWODNOŚCI W MIKROSYSTEMACH BEZPIECZEŃSTWA Artur Skrygulec, Andrzej Ruciński	101
KONWERSJA SYGNAŁÓW CYFROWYCH POMIĘDZY TELEKOMUNIKACYJNYMI I MULTIMEDIALNYMI SZYBKOŚCIAMI PRÓBKOWANIA Marek Blok	115
NOWY ESTYMATOR TONU KRTANIOWEGO Marek Blok, Mirosław Rojewski, Adam Sobociński	125
CYFROWY SYSTEM REJESTRACJI I REKONSTRUKCJI SYGNAŁU MOWY DLA POTRZEB LOTNICTWA WOJSKOWEGO Andrzej Czyżewski, Andrzej Kaczmarek, Józef Kotus, Arkadiusz Pawlik, Andrzej Rypulak, Paweł Żwan	135
KWADRATUROWY DDS Z UŁAMKOWO-OPÓŹNIAJĄCYM FILTREM O STRUKTURZE FLASH-FARROW Ewa Hermanowicz, Mirosław Rojewski	143
EFEKTYWNY ADAPTACYJNY ALGORYTM TRANSFORMATY FALKOWEJ DAUBECHIES 4 Michał Jacymirski, Piotr Lipiński	153
ALGORYTM NORMALIZACJI POZIOMÓW GŁOŚNOŚCI DZWIĘKU ZAREJESTROWANEGO W PLIKACH Przemysław Maziewski	161
MODELOWANIE PERFUZJI MÓZGU W BADANIACH MRI Renata Kalicka	169
INTERNETOWY SYSTEM DIAGNOZOWANIA ZMIAN MELANOCYTOWYCH SKÓRY Wiesław Paja	177

OPTIMALIZACJA POBUDZEŃ DLA CELÓW IDENTYFIKACJI PARAMETRÓW KOMPARTMENTOWYCH MODELI SYSTEMÓW FARMAKOKINETYCZNYCH Anna Pietrenko-Dąbrowska, Renata Kalicka	183
SYNTEZA OBRAZÓW PARAMETRYCZNYCH W BADANIU PERFUZJI MÓZGU METODĄ MRI Jacek Rumiński, Barbara Bobek-Billewicz	191
WEB SERVICES FRAMEWORK – STANDARDY, KORZYŚCI IMPLEMENTACJI ORAZ PRZYKŁADY ZASTOSOWAŃ Patryk Babiarz, Jacek Jakiela, Maciej Piotrowski, Bartosz Pomianek	199
ROZPROSZONE USŁUGI OBLICZENIOWE NA KLASTRACH TASK Z DOSTĘPEM PRZEZ WWW I „WEB SERVICES” Paweł Czarnul, Michał Bajor, Anna Banaszczyk, Paweł Buszkiewicz, Marcin Fiszer, Marcin Frączak, Michał Kławikowski, Jacek Rakiej, Katarzyna Ramczykowska, Krzysztof Suchcicki	207
WYKORZYSTANIE SERWERÓW UDDI DLA SYSTEMÓW ZDALNEJ EDUKACJI Madian dit Tiéman Diarra, Agnieszka Gwoździńska, Jerzy Kaczmarek	215
OBSZARY ZASTOSOWAŃ DYSTRYBUCJI CDLINUX.PL Jerzy Kaczmarek, Michał Wróbel	221
ZASTOSOWANIE ALGORYTMU HEURYSTYCZNEGO DO WYZNACZANIA KLAS UŻYTKOWNIKÓW KOOPERUJĄCYCH W SIECIOWYM SYSTEMIE INFORMATYCZNYM Radosław P. Katarzyniak	227
ZARZĄDZANIE USŁUGAMI W NOWOCZESNYCH SYSTEMACH WEBOWYCH Monika Koprowska, Rafał Sawzdargo	233
INTERNETOWY SYSTEM CZASU RZECZYWISTEGO DO AKWIZYCJI I WIZUALIZACJI OBRAZÓW RADAROWYCH Marek Moszyński, Jerzy Demkowicz, Andrzej Partyka	241
WARSTWA PREZENTACJI W WIELOWARSTWOWYCH SYSTEMACH INFORMACYJNYCH Bartosz Paliświat, Jerzy Nawrocki	249
MODEL ZAUFANIA DLA SYSTEMÓW WEBOWYCH Paweł Sachse, Krzysztof Juszczyzyn	257
TRENDY ROZWOJOWE TECHNOLOGII RADIA PROGRAMOWALNEGO Jacek Stefański	265
OPTIMALIZACJA PEWNYCH FUNKCJI KLASYFIKATORA BINARNEGO W PROCESIE POZYSKIWANIA INFORMACJI Z DANYCH WIELOKATEGORYJNYCH Mariusz Wrzesień	271
GSM CORDLESS TELEPHONY SYSTEM – PERSPEKTYWA ROZWOJU TRANSMISJI DANYCH W SYSTEMIE GSM Małgorzata Gajewska, Sławomir Gajewski	279
BADANIA ALGORYTMÓW WYBORU TRAS W NISKOORBITOWYCH SIECIACH SATELITARNYCH Janusz Jurski, Józef Woźniak	287
OCENA EFEKTYWNOŚCI PRACY STANDARDU BLUETOOTH Tomasz Klajbor, Józef Woźniak	295
METODY PRZYSPIESZENIA TRANSMISJI DANYCH W SYSTEMIE UMTS Krzysztof Małek, Dominik Rutkowski, Maciej Sosnowski	303
OCENA JAKOŚCI TRANSMISJI DANYCH W INTERFEJSIE RADIOWYM SYSTEMU UMTS Z WYKORZYSTANIEM DOSTĘPNYCH METOD KODOWANIA KANAŁOWEGO Andrzej Marczak, Rafał Niski	311
PERFORMANCE ANALYSIS AND OPTIMIZATION OF THE RADIO LINK CONTROL LAYER IN UMTS Paweł Matusz, Józef Woźniak	319

Andrzej Kraśniewski, Józef Lubacz

**Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska**

EWOLUCJA SZKOLNICTWA WYŻSZEGO W EUROPIE I JEJ KONSEKWENCJE DLA KSZTAŁCENIA W OBSZARZE TECHNIK I TECHNOLOGII INFORMACYJNYCH

Streszczenie

Prezentowano zasadnicze tendencje i oczekiwane kierunki dalszych zmian w ramach procesu harmonizowania systemów szkolnictwa wyższego w krajach europejskich, określanego jako Proces Boloński. Na tym tle sformułowano kilka tez dotyczących pożądanых cech systemu kształcenia studentów uczelni technicznych w szeroko rozumianej dziedzinie technik i technologii informacyjnych..

1. MIĘDZYKONKOWE UWARUNKOWANIA KSZTAŁCENIA INŻYNIERÓW

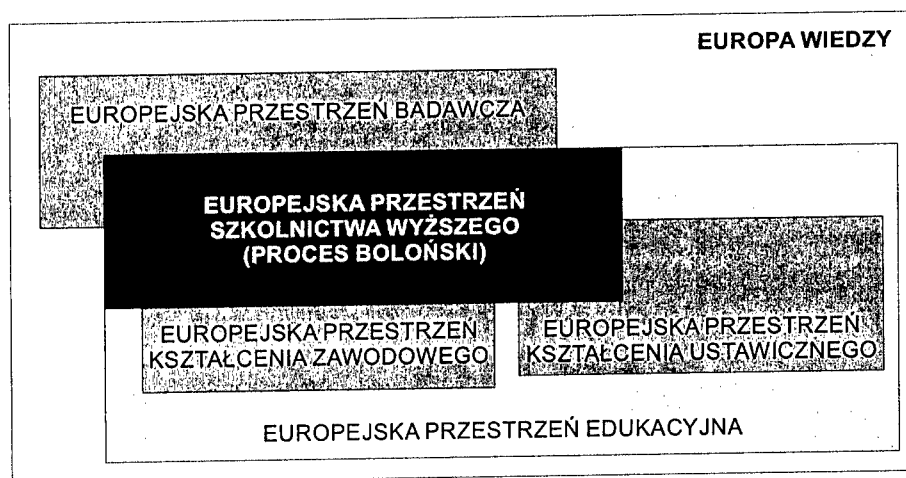
Kształcenie inżynierów w polskich uczelniach technicznych musi być dziś rozpatrywane w kontekście międzynarodowym – przede wszystkim europejskim. Wynika to co najmniej z dwóch powodów:

W warunkach globalnej gospodarki, absolwenci polskich uczelni będą coraz częściej pracować za granicą lub w działających w Polsce filiach zagranicznych lub międzynarodowych firm. Przystąpienie Polski do Unii Europejskiej zapewne nasili te procesy, ułatwiając swobodny przepływ wykształconych w Polsce inżynierów na rynki pracy innych krajów europejskich.

Od kilku lat realizowany jest – silnie wspierany przez Komisję Europejską - proces tworzenia Europejskiej Przestrzeni Szkolnictwa Wyższego (*European Higher Education Area*), określane często jako Proces Boloński. Jego zasięg i postęp jest już dziś na tyle znaczący, że ignorowanie zachodzących w ramach tego procesu zmian nie jest możliwe. Jakiegokolwiek przedsięwzięcia reformatorskie w obszarze szkolnictwa wyższego wyraźnie sprzeczne z założeniami Procesu Bolońskiego naraziłyby podejmujące je uczelnie na międzynarodową izolację, a absolwentom stworzyły mniej korzystne warunki startu zawodowego w jednoczącej się Europie.

Proces tworzenia Europejskiej Przestrzeni Szkolnictwa Wyższego (Proces Boloński) jest jednym z elementów szeroko zakrojonych działań zmierzających do utworzenia Europejskiej Przestrzeni Edukacyjnej (*European Area of Education and Training*) i jako taki pozostaje w związku z procesem tworzenia Europejskiej Przestrzeni Kształcenia Ustawicznego (*European Area of Lifelong Learning*) oraz z procesem tworzenia Europejskiej Przestrzeni Kształcenia Zawodowego, określanym jako *Bruges-Copenhagen Process*. Proces Boloński pozostaje także w silnym związku z procesem tworzenia Europejskiej Przestrzeni Badawczej (*European Research Area*). Oba te procesy są kluczowymi elementami procesu mającego doprowadzić do realizacji stworzonej przez polityków wizji Europy Wiedzy (*Europe of Knowledge*), stanowiącej kluczowy element tzw. Strategii Lizbońskiej (rys. 1).

„Posadowienie” Procesu Bolońskiego w kontekście tworzenia Europy Wiedzy powoduje, że kształcenie w uczelniach technicznych nabrało większego znaczenia (w porównaniu z kształceniem w innych obszarach) i stało się przedmiotem zainteresowania szerszego grona osób, w tym polityków odpowiedzialnych za strategię rozwoju Europy i poszczególnych krajów. W dokumencie Komisji Europejskiej *Education and training in Europe: The work programme on the future objectives of education and training in Europe* sformułowano 13 zadań dla europejskiego systemu edukacji; wśród tych zadań znalazło się zwiększenie naboru na studia w dziedzinie nauk ścisłych i studia techniczne [CEC02]. W dokumencie Komisji Europejskiej *The role of the universities in the Europe of knowledge* wśród wielu postawionych pytań znalazło się i takie: Co należy zrobić, aby studia na kierunkach ścisłych i technicznych, a także kariery zawodowe w tych dziedzinach stały się bardziej atrakcyjne? [CEC03].



Rys. 1. Proces Boloński jako element kształtowania Europy Wiedzy

W ocenie ekspertów, realizacja Strategii Lizbońskiej, a ściślej sformułowanego przez szefów rządów krajów Unii Europejskiej postulatu zwiększenia do 2010 r. nakładów na prace badawczo-rozwojowe do wymiaru 3% PKB będzie wymagała utworzenia w Europie ok. 500000 nowych miejsc pracy związanych z badaniami naukowymi, z czego znaczna część związana byłaby z badaniami w obszarze nauk ścisłych i nowoczesnych technologii. Tworzenie Europy Wiedzy stanowi więc wielką szansę dla absolwentów uczelni technicz-

nych, ale... absolwentów właściwie przygotowanych. Właściwe przygotowanie do pracy w sektorze badawczo-rozwojowym oznacza m.in. umiejętność myślenia abstrakcyjnego i systemowego oraz umiejętność samodzielnego stawiania i rozwiązywania problemów. Z pewnością nie służy takiemu przygotowaniu model kształcenia, w którym dominuje przekazywanie wiedzy i umiejętności z zakresu wąsko pojętej specjalności.

Różne przesłanki związane ze zmianami zachodzącymi w systemach szkolnictwa wyższego w krajach europejskich skłaniają do formułowania wielu innych postulatów dotyczących modelu kształcenia na uczelniach technicznych. Celem artykułu jest zatem omówienie procesów zachodzących w europejskim szkolnictwie wyższym i na tym tle przedstawienie kilku też dotyczących pożądaných cech systemu kształcenia studentów uczelni technicznych w szeroko rozumianym obszarze technik i technologii informacyjnych, obejmującym także telekomunikację.

2. ZMIANY W EUROPEJSKIM SZKOLNICTWIE WYŻSZYM – PROCES BOŁOŃSKI

Deklaracja Bolońska, podpisana 19 czerwca 1999 r. przez ministrów odpowiedzialnych za szkolnictwo wyższe w 29 krajach europejskich, zapoczątkowała proces istotnych zmian w systemach edukacji poszczególnych państw. Proces ten, nazywany często Procesem Bolońskim, zmierza do utworzenia do roku 2010 – w wyniku uzgodnienia pewnych ogólnych zasad organizacji kształcenia – Europejskiej Przestrzeni Szkolnictwa Wyższego (*European Higher Education Area*).

Proces Boloński stanowi próbę wypracowania wspólnej „europejskiej” reakcji na problemy występujące w większości krajów, tak aby:

- stworzyć warunki do mobilności obywateli,
- dostosować system kształcenia do potrzeb rynku pracy, a zwłaszcza doprowadzić do poprawy „zatrudnialności”,
- podnieść atrakcyjność i poprawić pozycję konkurencyjną systemu szkolnictwa wyższego w Europie, tak aby odpowiadała ona wkładowi tego obszaru w rozwój cywilizacji.

Celem zachodzących procesów integracyjnych nie jest standaryzacja, lecz raczej „harmonizacja”, tzn. wypracowanie zasad współdziałania, z uwzględnieniem zróżnicowania i autonomii poszczególnych państw i uczelni.

Z formalnego punktu widzenia najistotniejszymi dokumentami określającymi charakter Procesu Bolońskiego są deklaracje i komunikaty sygnowane przez ministrów odpowiedzialnych za szkolnictwo wyższe w krajach europejskich. Pierwszym dokumentem tego typu, poprzedzającym Deklarację Bolońską, była Deklaracja Sorbońska z maja 1998 r., pod którą podpisy złożyli ministrowie Francji, Niemiec, W. Brytanii i Włoch. Zawarta w Deklaracji Sorbońskiej idea „harmonizacji” struktury systemów szkolnictwa wyższego w celu zwiększenia mobilności i poprawy „zatrudnialności” została następnie rozwinięta w Deklaracji Bolońskiej, podpisanej przez ministrów 29 krajów (w tym Polski), następnie w Komunikacie Praskim z maja 2001 r., a ostatnio w Komunikacie Berlińskim z września 2003 r. Liczba państw zaangażowanych w Proces Boloński systematycznie wzrasta; w wyniku decyzji podjętych w Berlinie w tworzeniu Europejskiej Przestrzeni Szkolnictwa Wyższego uczestniczy obecnie formalnie 40 krajów.

Proces Boloński ma bogatą „literaturę”. Wśród ostatnio opublikowanych materiałów zwraca uwagę przygotowywany na zamówienie Komisji Europejskiej raport *Trends in Learning Structures in Higher Education (Trends III)* [TrendsIII]. Źródłem aktualnych

informacji o przebiegu Procesu Bolońskiego, zawierającym podstawowe dokumenty oraz wiele innych materiałów, jest witryna internetowa <http://www.bologna-bergen2005.de>. Rozszerzenie zagadnień omówionych w niniejszym artykule można znaleźć w [Kras03].

2.1. Podstawowe postulaty Procesu Bolońskiego

W Deklaracji Bolońskiej zawarte jest sześć postulatów wskazujących sposoby realizacji celów przyświecających idei tworzenia Europejskiej Przestrzeni Szkolnictwa Wyższego:

- wprowadzenie systemu „łatwo czytelnych” i porównywalnych stopni (dyplomów),
- wprowadzenie studiów dwustopniowych,
- wprowadzenie punktowego systemu rozliczania osiągnięć studentów (ECTS),
- usuwanie przeszkód ograniczających mobilność studentów i pracowników,
- współdziałanie w zakresie zapewniania jakości kształcenia,
- propagowanie problematyki europejskiej w kształceniu.

W Komunikacie Praskim wskazano kolejne istotne elementy Europejskiej Przestrzeni Szkolnictwa Wyższego:

- kształcenie ustawiczne,
- współdziałanie uczelni i studentów w realizacji Procesu Bolońskiego,
- propagowanie atrakcyjności Europejskiej Przestrzeni Szkolnictwa Wyższego poza Europą.

Komunikat Berliński wskazuje nowe aspekty Procesu Bolońskiego, podkreślając:

- związek kształcenia i badań naukowych oraz znaczenie badań jako integralnej części szkolnictwa wyższego,
- potrzebę rozszerzenia dwustopniowego systemu studiów (zdefiniowanego w Deklaracji Bolońskiej) o studia III stopnia – studia doktoranckie,
- potrzebę kształcenia interdyscyplinarnego.

Poniżej omówiono sposoby realizacji niektórych z ww. postulatów Procesu Bolońskiego – szczególnie istotnych z punktu widzenia kształcenia w uczelniach technicznych.

2.2. Studia dwustopniowe

Wprowadzanie studiów dwustopniowych w uczelniach europejskich przebiega dość szybko; 53% uczelni wprowadziło lub właśnie wprowadza ten system, a kolejne 36% uczelni zamierza to zrobić w najbliższej przyszłości [TrendsIII].

Model studiów dwustopniowych jest obecnie dość powszechnie rozumiany w sposób następujący (używane poniżej terminy *bachelor* i *master*, pisane małą literą, rozumiane są jako niezależne od kraju, ogólne nazwy stopni i dyplomów odpowiadających ukończeniu studiów I i II stopnia):

- uzyskanie dyplomu *bachelor* (ukończenie studiów I stopnia) wymaga zdobycia 180-240 punktów ECTS,

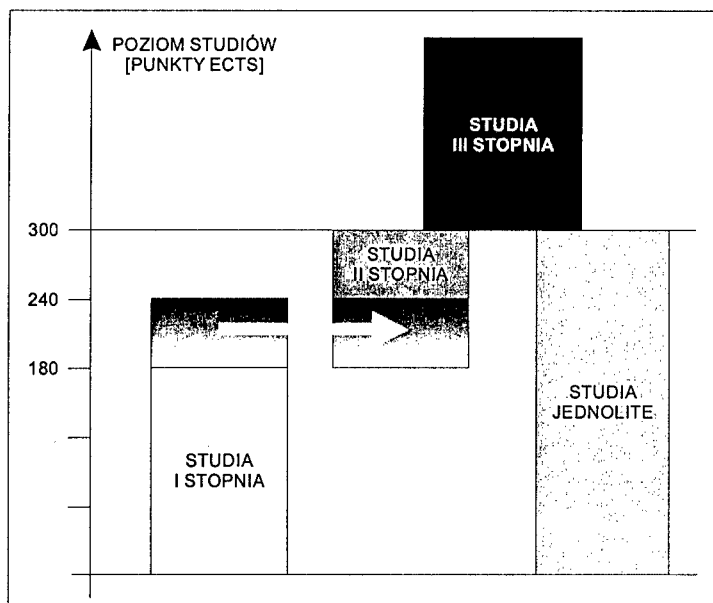
- uzyskanie dyplomu *master* (ukończenie studiów II stopnia) wymaga zdobycia 300 punktów ECTS, licząc od początku studiów, z czego co najmniej 60 punktów ECTS musi być uzyskane na poziomie „graduate” w obszarze specjalności uwi-docznionym na dyplomie.

Przyjmuje się zatem, że ścieżka kształcenia prowadząca do uzyskania dyplomu *master* może mieć w szczególności następującą postać (symbole B i M oznaczają liczbę punktów ECTS uzyskiwaną na studiach prowadzących do dyplomu *bachelor* i *master*) [TrendsIII]:

- 180 B + 120 M,
- 240 B + 90-120 M, z czego 30-60 M może być uzyskane w wyniku uznania osiągnięć z ostatniego roku studiów I stopnia,
- 300 M, co oznacza zintegrowany program studiów I i II stopnia, prowadzący bezpośrednio do dyplomu *master*.

Należy jednak podkreślić, że model kształcenia odpowiadający jednolitym studiom magisterskim – choć dopuszczalny – jest traktowany jako rozwiązanie nietypowe. W szczególności, uczelnie realizujące kształcenie w tym trybie mają znacznie utrudnione warunki ścisłej współpracy w obszarze edukacji z uczelniami z innych krajów, zaś studenci realizujący kształcenie w tym trybie pozbawieni są możliwości korzystania z różnych wspieranych przez Komisję Europejską form mobilności.

Decyzją ministrów zebranych na Szczycie Berlińskim tak rozumiany model studiów dwustopniowych został rozszerzony o studia doktoranckie, traktowane jako studia III stopnia (rys. 2).

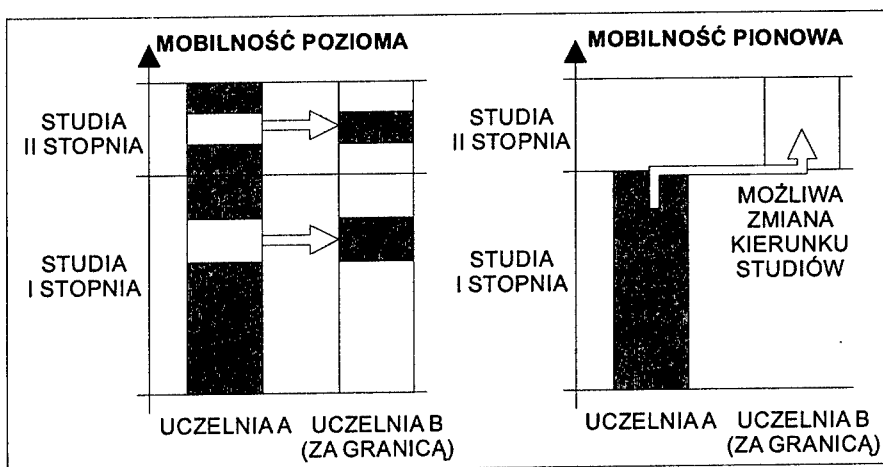


Rys. 2. Struktura systemu studiów

2.3. Usuwanie przeszkód ograniczających mobilność studentów i pracowników

Mobilności studentów sprzyjają specjalne programy finansowane bezpośrednio lub pośrednio przez Komisję Europejską. W ramach programu Socrates-Erasmus w latach 1997-2002 ponad 1 mln studentów miało okazję zrealizować część programu studiów w trakcie trwającego zazwyczaj kilka miesięcy pobytu w uczelni zagranicznej. Oznacza to, że ok. 5% obecnych absolwentów uczelni europejskich ma doświadczenie międzynarodowe związane ze studiami odbywanymi – przynajmniej częściowo – za granicą.

Realizacja części programu studiów I lub II stopnia w innej uczelni, w kraju lub za granicą, określana jest jako mobilność pozioma (*horizontal mobility*) [TrendsIII]. Opcją coraz częściej rozpatrywaną podczas projektowania indywidualnej ścieżki kształcenia staje się inna forma mobilności – mobilność pionowa (*vertical mobility*) [TrendsIII], oznaczająca zmianę uczelni po ukończeniu studiów I stopnia, często połączoną ze zmianą kierunku studiów (rys. 3).



Rys. 3. Mobilność studentów

Najbardziej rozwinięta forma mobilności (poziomej) jest związana z programami studiów prowadzonymi wspólnie przez uczelnie z różnych krajów (*joint degrees*). Program taki realizowany jest na podstawie wieloletniej umowy dwóch lub większej liczby uczelni, a jego cechami są:

- wspólnie opracowane plany studiów i programy nauczania,
- porównywalne okresy studiowania w uczelniach partnerskich,
- wspólnie prowadzone prace dyplomowe i egzaminy dyplomowe,
- wymiana wykładowców między uczelniami partnerskimi,
- „wspólny dyplom”.

W celu przynajmniej częściowej likwidacji barier hamujących rozwój programów studiów prowadzonych wspólnie przez uczelnie z różnych krajów na Szczycie Berlińskim ministrowie podjęli zobowiązanie wprowadzenia do 2005 r. odpowiednich regulacji prawnych dotyczących mechanizmów tworzenia, uznawalności i akredytacji tego rodzaju

studiów, a także wydawania „wspólnych dyplomów” – dokumentów sygnowanych przez przedstawicieli wszystkich uczelni uczestniczących w realizacji programu.

Programy studiów prowadzone wspólnie przez uczelnie z różnych krajów mają szansę stać się „znakiem firmowym” europejskiego szkolnictwa wyższego; są już dziś „ukochanym dzieckiem” Komisji Europejskiej. W opinii autorów raportu *Trends III* rozwój wspólnych programów studiów jest na tyle istotny, że nie wspierając – także finansowo – tego rozwoju, poszczególne kraje oraz poszczególne uczelnie tracą doskonałą sposobność wypracowania sobie korzystnej pozycji w Europejskiej Przestrzeni Szkolnictwa Wyższego [TrendsIII].

2.4. Wprowadzenie systemu „łatwo czytelnych” i porównywalnych stopni (dyplomów)

Podstawowym mechanizmem prowadzącym do poprawy „czytelności” stopni (dyplomów) jest suplement do dyplomu (*Diploma Supplement*). Suplement do dyplomu zawiera informacje niezbędne do określenia poziomu i charakteru wykształcenia uzyskanego przez absolwenta studiów wyższych:

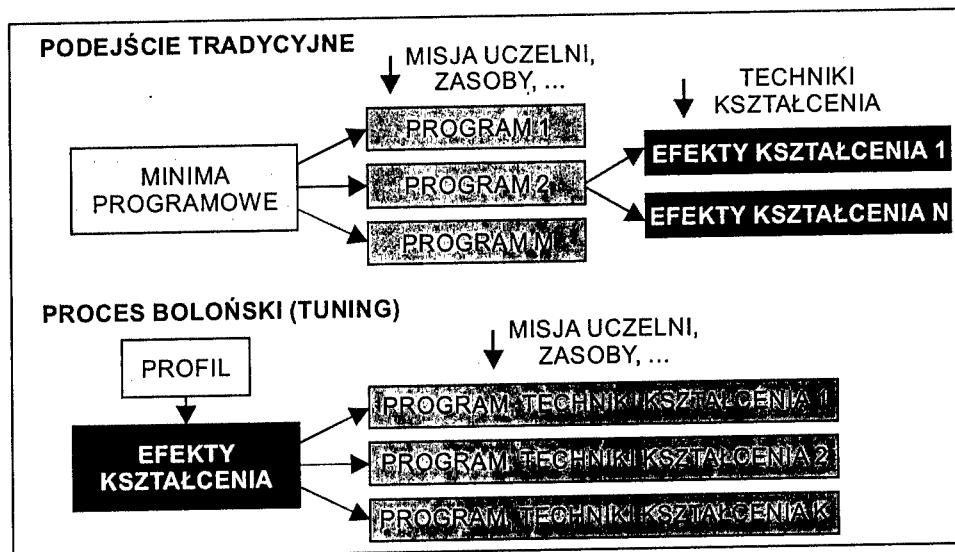
- charakterystykę systemu studiów wyższych w kraju, w którym prowadzone są studia,
- charakterystykę zrealizowanego programu studiów,
- rejestr indywidualnych osiągnięć studenta.

Suplement do dyplomu jest dokumentem wydawanym w języku kraju, w którym prowadzone są studia, oraz w jednym lub większej liczbie „języków międzynarodowych”. Na Szczycie Berlińskim ustalono, że począwszy od roku 2005 w krajach uczestniczących w Procesie Bolońskim suplement do dyplomu będzie wydawany każdemu absolwentowi studiów wyższych, bez dodatkowej opłaty.

Jednym z istotnych zadań w Procesie Bolońskim jest możliwie jak najszybsze zdefiniowanie struktury stopni (tytułów) zawodowych (*qualification framework*) w poszczególnych krajach, oparte na opisie „wyjścia” (*outcome based qualification framework*). Każdy z elementów tej struktury (każdy typ dyplomu) powinien być scharakteryzowany przez [Reding03]:

- poziom, mierzony nakładem pracy studenta wyrażonym w punktach ECTS,
- profil, określający charakter uzyskanych kwalifikacji
- efekty kształcenia (*learning outcomes, competences*), określające zakres wiedzy i umiejętności (*knowledge, skills*) posiadanych przez absolwenta.

Realizacja tak postawionego zadania wymaga nowego podejścia do tworzenia programów studiów. Nowa metodyka tworzenia programów studiów, zaproponowana w ramach projektu pilotażowego *Tuning Educational Structures in Europe*, finansowanego ze środków Komisji Europejskiej, zakłada, że punktem wyjścia są efekty kształcenia. Na tej podstawie każda uczelnia tworzy swój „autorski” program studiów i dobiera odpowiednie techniki nauczania [Tuning03]. Programy opracowane w różnych uczelniach mogą się różnić bardzo znacznie, lecz ich realizacja prowadzi do osiągnięcia podobnych celów, określonych przez definicję „wyjścia”, tzn. efektów kształcenia. Takie podejście z jednej strony stwarza szerokie możliwości eksperymentowania i wdrażania nowatorskich koncepcji dydaktycznych, a z drugiej strony – gwarantuje pożądany efekt końcowy, tzn. właściwe kwalifikacje absolwenta. Różnicę między tradycyjną metodyką opracowywania programów studiów a podejściem zaproponowanym w ramach projektu *Tuning* zilustrowano na rys. 4.



Rys. 4. Definiowanie programów kształcenia

2.5. Proces Boloński w ocenie środowiska „inżynierów”

Z punktu widzenia uczelni technicznych istotne jest stanowisko, jakie wobec Procesu Bolońskiego zajmują międzynarodowe organizacje zaangażowane w sprawy kształcenia inżynierów. Ważnym dokumentem wyrażającym takie stanowisko jest ogłoszony w lutym 2003 r. wspólny komunikat CESAER (*Conference of European Schools for Advanced Engineering Education and Research*) i SEFI (*Société Européenne pour la Formation des Ingénieurs - European Society for Engineering Education*) [SEFI03]. Przy opracowaniu tego dokumentu wzięto również pod uwagę stanowisko innych organizacji zajmujących się kształceniem w uczelniach technicznych, m.in. BEST, FEANI, IAESTE i E4.

We wspólnym komunikacie CESAER i SEFI wyrażają zdecydowane poparcie dla idei Europejskiej Przestrzeni Szkolnictwa Wyższego, stwierdzając jednocześnie, że – ze względu na tradycję i charakter kształcenia inżynierów – postulaty Deklaracji Bolońskiej powinny być realizowane w odniesieniu do studiów technicznych w nieco inny sposób niż to ma miejsce w przypadku innych obszarów studiów. Wynikają stąd m.in. następujące zalecenia dotyczące dalszych etapów Procesu Bolońskiego:

- Oprócz upowszechnienia studiów dwustopniowych zaleca się utrzymanie zintegrowanych programów studiów prowadzących bezpośrednio do dyplomu *master* (jednolitych studiów magisterskich). Postulat ten budzi jednak kontrowersje. Zdaniem V. Reding, pełniącej funkcję komisarza Unii Europejskiej ds. edukacji i kultury, współlistnienie studiów I i II stopnia oraz zintegrowanych studiów prowadzących bezpośrednio do dyplomu *master* może doprowadzić do komplikacji w strukturze systemu kształcenia [Reding03].
- Studia techniczne o profilu akademickim powinny kończyć się co najmniej na poziomie *master*. Zgodnie z tym postulatem studia I stopnia o profilu akademickim powinny być więc traktowane przede wszystkim jako przygotowanie do studiów II stopnia (niekoniecznie na tym samym „kierunku”).

- Niezależnie od zróżnicowania studiów I stopnia, a także studiów II stopnia, system kształcenia musi zachować wewnętrzną drożność, tzn. absolwent studiów I stopnia bez względu na ich charakter musi mieć sensowną możliwość kontynuowania kształcenia. W szczególności, struktura systemu studiów powinna umożliwiać kontynuowanie kształcenia absolwentom studiów I stopnia o profilu zawodowym.
- Definicje „dyplomów” powinny być oparte na efektach kształcenia.

3. PROCES BOŁOŃSKI A MODEL STUDIÓW W POLSKICH UCZELNIACH TECHNICZNYCH

Dość powszechne wprowadzanie w polskich uczelniach technicznych systemu studiów dwustopniowych nie jest, jak sądzą niektórzy, „uleganiem dyktatowi bolońskiemu”, lecz wynika z potrzeby dostosowania systemu kształcenia do nowych uwarunkowań cywilizacyjnych, w których bez wykształcenia na poziomie wyższym coraz trudniej o pracę. Odsetek populacji kształcanej na tym poziomie musi być dziś znacznie większy niż w przeszłości. Uzyskanie określonych kwalifikacji nie będzie jednak gwarantować stabilności zatrudnienia, a w związku z tym konieczność częstych zmian kwalifikacji stanie się normą. Wynika stąd, że system jednolitych studiów magisterskich, ukształtowany w innych warunkach (kształcenie małego odsetka najzdolniejszej młodzieży, większa trwałość raz zdobytych kwalifikacji, stabilność zatrudnienia i rodzaju pracy, itd.), nie jest dostosowany do nowych potrzeb. Nie znaczy to jednak, że w pewnych dziedzinach (np. w medycynie) nie powinno się zachować tradycyjnych jednolitych studiów magisterskich, ale liczba takich dziedzin jest stosunkowo mała. Nie ma też powodu, by przejście na system studiów dwustopniowych miało prowadzić do zaniku kształcenia elitarnego.

Aby wprowadzenie systemu studiów dwustopniowych mogło przynieść spodziewane korzyści, musi mu towarzyszyć istotna zmiana założeń programowych w stosunku do tradycyjnych studiów magisterskich. Studia pierwszego stopnia nie mogą być „przyciętymi”, czy „strywalizowanymi” studiami magisterskimi. Sensownych studiów dwustopniowych nie da się też stworzyć poprzez proste podzielenie programów tradycyjnych studiów magisterskich na dwie części, bowiem inne powinny być zasadnicze cele kształcenia na studiach pierwszego i drugiego stopnia. Nieporozumieniem jest też następujące, pozornie „logiczne” rozumowanie: skoro mamy dać wykształcenie zawodowe w czasie krótszym niż na tradycyjnych studiach magisterskich, to musi ono być skupione wokół węższej specjalności. A zatem przejście na dwustopniowy system studiów wymaga gruntownej przebudowy programów nauczania, przy dobrze określonych celach i profilach kształcenia na obu stopniach. W tym jest sedno sprawy, a zarazem główna trudność, której pokonanie wymaga dużego wysiłku merytorycznego i organizacyjnego.

Zgodnie z obowiązującym ustawodawstwem, dyskutując na temat organizacji systemu studiów w polskich uczelniach należałoby używać terminów „studia zawodowe” i „uzupełniające studia magisterskie”. Są to określenia niefortunne, prowadzące do nieporozumień. W niniejszym tekście konsekwentnie są stosowane terminy „studia I stopnia” (licencjackie lub inżynierskie) i „studia II stopnia” (studia magisterskie), które są pojętniejsze, budzą mniej kontrowersji i są zgodne z nazewnictwem używanym w dokumentach dotyczących Europejskiej Przestrzeni Szkolnictwa Wyższego.

Wydaje się, że podstawową wątpliwość związaną z wprowadzaniem studiów dwustopniowych, między innymi w uczelniach technicznych, można sprowadzić do pytania: Czy celem studiów I stopnia powinno być przygotowanie do wykonywania określonego zawodu? Można na nie odpowiedzieć twierdząco, pod warunkiem jednak, iż inaczej

niż nakazuje to tradycja, interpretuje się określenie „przygotowanie do wykonywania zawodu”. Rzecz w tym, że tradycyjne pojmowanie terminu „wykształcenie zawodowe” wiązało się z nabyciem takich umiejętności, które można było eksploatować – bez zasadniczego ich uaktualnienia, czy modyfikacji – przez długi czas, nierzadko przez całe życie. Dzisiaj, jak już wspomniano, mamy do czynienia z radykalnie zmieniającą się sytuacją, w której dawna stabilność wykształcenia i związanego z nim zatrudnienia staje się coraz rzadsza. Zaplanowanie „jednorazowego” wykształcenia zawodowego, które nie traciłoby aktualności i przydatności na rynku pracy przez długi czas, jest dzisiaj praktycznie niemożliwe. Wobec tego, także cel kształcenia, które chcielibyśmy określać jako zawodowe, musi być inny, niż jeszcze nie tak dawno temu. Współczesne kształcenie zawodowe w ramach studiów I stopnia może i powinno być ukierunkowane na umożliwienie absolwentowi podjęcia pracy, z reguły pierwszej, bądź dalszego kształcenia się w ramach szeroko określonego profilu zawodowego, a nie na przygotowanie absolwenta do wykonywania ściśle określonego zawodu. Inaczej mówiąc, powinno to być kształcenie prowadzące do uzyskania „przepustki” do podjęcia pracy w szeroko określonej dziedzinie, a nie do uzyskania wąsko określonych umiejętności czy uprawnień zawodowych, a także przygotowywać do podjęcia dalszej nauki – samodzielnej, organizowanej przez pracodawcę, czy na uczelni w ramach studiów wyższego stopnia bądź podyplomowych.

Wobec tego, w ramach studiów I stopnia powinno się propagować tworzenie możliwie szerokich i wszechstronnych programów nauczania, w szczególności mających cechy interdyscyplinarności, a nie programów nastawionych na wąskie specjalności zawodowe. Tak rozumiane kształcenie można by nazwać makrokierunkowym czy międzykierunkowym, jeśli za punkt odniesienia przyjąć obecnie obowiązujące w Polsce kierunki kształcenia (określenia te występują w projekcie nowego prawa o szkolnictwie wyższym [Ustawa04]).

Kształtując programy studiów II stopnia należy pamiętać o tym, że jednym z ważnych celów wprowadzania studiów dwustopniowych jest udostępnienie drugiego stopnia studiów dla absolwentów studiów I stopnia spoza macierzystej jednostki, tak aby obok „mobilności poziomej” możliwa była realizacja „mobilności pionowej” (patrz rys. 3). Studia II stopnia powinny być nie tylko drugim etapem studiów dla „własnych” studentów, którzy chcą poszerzyć swą wiedzę, bądź zmodyfikować profil specjalizacji uzyskany na studiach I stopnia, ale być także otwarte dla osób, które uzyskały dyplom licencjata lub inżyniera w pokrewnej dziedzinie wiedzy na innym wydziale/uczelni, z których część może chcieć powrócić na uczelnię po okresie pracy zawodowej. System studiów II stopnia powinien być zatem elastyczny, w tym sensie, aby umożliwiał indywidualizowanie programów nauczania – ich dostosowywanie do zróżnicowanego przygotowania merytorycznego studentów przyjmowanych na te studia. W praktyce, program studiów dla studentów spoza macierzystej jednostki będzie często musiał być rozszerzony o przedmioty „wyrównujące” – zazwyczaj z programu studiów I stopnia, a więc dla tych studentów czas trwania studiów będzie z reguły dłuższy.

Kwestia modelu struktury studiów (nie tylko technicznych) i zasadniczych koncepcji programowych zbliżających nas do rozwiązań europejskich powinna być tematem powszechnej debaty i doprowadzić do ustaleń o charakterze ogólnokrajowym. Należy jednak podkreślić, że niezależnie od przedsięwzięć na poziomie kraju, poszczególne uczelnie, a nawet ich pojedyncze jednostki, mogą zrobić wiele „na własnym podwórku”, aby wypracować sobie jak najlepszą pozycję w tworzącej się Europejskiej Przestrzeni Szkolnictwa Wyższego. Konieczne jest w szczególności konsekwentne wprowadzanie

oferty edukacyjnej w języku angielskim, programów studiów realizowanych wspólnie z uczelniami z innych krajów, itd.

4. KSZTAŁCENIE W OBSZARZE TECHNIK I TECHNOLOGII INFORMACYJNYCH

Po rozważaniach dotyczących ogólnych zagadnień kształcenia w uczelniach technicznych przyjrzyjmy się bliżej pożądanym cechom systemu kształcenia w odniesieniu do obszaru wiedzy i umiejętności, który określono tu jako nauki, techniki i technologie informacyjne. Określenie to wymaga komentarza, zaczniemy więc od tej sprawy.

W ubiegłej dekadzie upowszechnił się angielskojęzyczne terminy *information science* oraz *information technology*. Są to terminy mające stosunkowo krótką historię, a ich zakres znaczeniowy nie jest jednoznaczny i podlega ewolucji. *Information science* – nauki informacyjne – jest terminem, który odnosi się do szeroko rozumianej wiedzy dotyczącej modelowania procesów informacyjnych zachodzących zarówno w tworach sztucznych („komputerach”), jak i w naturze. A więc *information science* ma inny zakres znaczeniowy niż termin informatyka, do którego przywykliśmy w Polsce, a który zasadniczo odpowiada angielskiemu *computer science & engineering*, ale z czasem stał się dość rozmyty – bywa interpretowany w sposób zawężony, jako *computer science* (bez członu *computer engineering*), ale także w sposób rozszerzony, obejmujący zarówno *computer science & engineering*, jaki i nauki informacyjne w szerszym, „pozakomputerowym” sensie. Pojawiają się oczywiście też różnorakie interpretacje pośrednie, a co więcej, różnie bywają rozumiane same terminy *computer science* oraz *computer engineering*. Warto w tym kontekście wziąć pod uwagę, że informatyka jako dyscyplina naukowa jest zaliczana, zgodnie z obowiązującymi w Polsce przepisami, do dwóch dziedzin naukowych: do nauk technicznych i do nauk matematycznych (jest to jedyny taki przypadek).

Należałoby dążyć do tego, aby termin informatyka był używany zgodnie z jego pierwotnym znaczeniem (*computer science & engineering*) oraz upowszechnić termin „nauki informacyjne”. Alternatywą jest rozszerzenie zakresu znaczeniowego terminu informatyka tak, aby obejmował także nauki informacyjne. Nie byłoby to jednak rozwiązanie rozsądne, gdyż spowodowałoby jeszcze większy chaos interpretacyjny, a poza tym pogłębiałoby obecną niezgodność terminologii polsko- i angielskojęzycznej. Już w obecnym stanie rzeczy, termin informatyka jest stosowany w sposób dość dowolny i prowadzący do nieporozumień, w szczególności w nazewnictwie obszarów kształcenia. Stosuje się go w politechnikach, uniwersytetach i uczelniach ekonomicznych, ale oczywiście programy nauczania ukryte w tych uczelniach pod nazwą informatyka różnią się na tyle istotnie, że określenie ich wspólnego jądra, na przykład w postaci wspólnych minimów programowych, jest praktycznie niemożliwe. Jeśli nazwy obszarów/kierunków kształcenia mają być adekwatne do ich zawartości programowej, a powinny, chociażby dlatego, aby nie wprowadzały w błąd kandydatów na studia (nie mówiąc już o celowym wprowadzaniu w błąd, by zwabić „klientów”), to zachodzi potrzeba zróżnicowania ich nazw. Celowi temu powinno służyć upowszechnienie terminu nauki informacyjne oraz uporządkowanie interpretacji i zastosowania terminu informatyka.

Rozpatrując tę kwestię, należy uwzględnić powiązanie powyższych terminów z coraz powszechniejszym terminem *information technology* (IT). Na polski należałoby go tłumaczyć jako techniki informacyjne, ale coraz częściej tłumaczy się go jako technologie informacyjne. Dzieje się tak głównie za sprawą mediów masowych, które zaczęły tłumaczyć angielskie *technology* na „technologie”, zamiast „techniki”. Wziąwszy pod

uwagę siłą „przekonywania” mediów masowych, sprawa zamiany pierwotnego znaczenia terminów „technika” i „technologia” wydaje się przesądzona – zapewne przyjdzie się do niej przyzwyczaić, a w końcu usankcjonować. W tytule niniejszego tekstu celowo zastosowano nieco nadmiarowe określenie „techniki i technologie informacyjne”, mając nadzieję, że pomoże ono uniknąć ewentualnych nieporozumień co do przedmiotu rozważań, które mogą się pojawić wśród tych, którzy jeszcze nie pogodzili się z używaniem terminu technologia w takim znaczeniu, jak w języku angielskim.

Ale poza kwestią językową, jest też problem zakresu znaczeniowego *information technology*. Terminem tym określa się technologie, które są związane ze wszystkimi aspektami operowania informacją – jej wytwarzaniem, przetwarzaniem, gromadzeniem, wyszukiwaniem, przesyłaniem, komunikowaniem, itp. (Równocześnie z *information technology* upowszechniany był termin *information-communication technology (ICT)*, o zasadniczo tym samym zakresie znaczeniowym, używany zamiennie; wydaje się, że termin ten z czasem straci popularność, gdyż jest de facto redundantny). *Information technologies* – technologie informacyjne – to zatem termin niezwykle pojemny, który swym zakresem znaczeniowym obejmuje obszary wiedzy i umiejętności tradycyjnie wiązane z informatyką i telekomunikacją, ale także z automatyką i cybernetyką, czy też z szeroko pojętą elektroniką (fotoniką, nanotechnologią, ...), w zakresie odnoszącym się w szczególności do warstwy fizycznej systemów operujących informacją. We wszystkich tych obszarach zagadnienia związane z abstrakcyjnym, „poza komputerowym” modelowaniem procesów informacyjnych mają coraz większe znaczenie, w związku z czym nauki informacyjne i technologie informacyjne pokrywają się w niemałym stopniu w sensie pojęciowym i metodologicznym.

W ostatnich latach pojawił się jeszcze inny termin, propagowany szczególnie w kontekście programów badawczo-rozwojowych Unii Europejskiej: *information society technologies (IST)*, to jest techniki/technologie społeczeństwa informacyjnego. Choć jest to określenie ukierunkowane głównie na podkreślenie aspektów aplikacyjnych technologii informacyjnych – na ich zastosowania we współczesnych społecznościach – to jednak stopniowo zaczęło być ono stosowane zamiennie z terminem techniki/technologie informacyjne. W dalszym ciągu niniejszego tekstu określenie to nie będzie stosowane, gdyż ma, na gust autorów, posmak nazbyt „socjotechniczny”.

Wobec przedstawionych okoliczności, godząc się z anglosaskim znaczeniem terminu technologia, w dalszej części tekstu tytułowe określenie „nauki, techniki i technologie informacyjne” zostanie skrócone do „nauki i technologie informacyjne” – NTI. Po tych uwagach i ustaleniach terminologicznych zajmijmy się bliższym scharakteryzowaniem NTI w kontekście ogólnych rozważań dotyczących problemów kształcenia, które zostały poruszone w poprzednim rozdziale.

W obowiązującej w Polsce klasyfikacji dyscyplin naukowych znajdują się: Automatyka i robotyka, Elektronika, Informatyka i Telekomunikacja, które bez wątpienia można zaliczyć do NTI. Z kolei na liście ministerialnych kierunków studiów znajdują się: Automatyka i robotyka, Elektronika i telekomunikacja oraz Informatyka (przed kilkunastu laty Elektronika i Telekomunikacja były osobnymi kierunkami studiów; inicjatywa ich ponownego rozdzielenia, z końca lat dziewięćdziesiątych, nie powiodła się z powodów pozamerytorycznych). Mamy więc do czynienia ze znacznym stopniem zgodności kierunków studiów z dyscyplinami naukowych. Zatem obszar NTI jest spójny pod względem zgodności obszaru edukacyjnego z naukowym, a ten charakteryzuje się znacznym stopniem powiązania merytorycznego składowych dyscyplin naukowych, jako związanych z różnymi aspektami operowania informacją. Stwarza to dobrą podstawę do tworzenia spójnych

międzykierunkowych i makrokierunkowych programów nauczania, które – zgodnie z wcześniejszymi uwagami – powinny odgrywać istotną rolę przy konsekwentnym wprowadzaniu studiów dwustopniowych.

Do idei kształcenia międzykierunkowego i makrokierunkowego podnosi się często zastrzeżenie, że wnosi ona ryzyko amorficzności programów nauczania – braku skoncentrowania na „dobrze określonych” obszarach wiedzy. Obawy takie mogą być uzasadnione w stosunku do takich programów, które są związane z odległymi dyscyplinami naukowymi, a szczególnie takimi, które należą do innych dziedzin nauki. Nie powinno to jednak dotyczyć kształcenia w obszarze NTI, które związane jest z bliskimi sobie dyscyplinami naukowych, wchodzącymi w skład nauk technicznych (szczególną pozycję zajmuje informatyka, która, jak już wspomniano, jest zaliczana także do dziedziny nauk matematycznych). Zastrzeżenia trudno byłoby także zgłosić z punktu widzenia wykładni zasad regulujących umieszczanie kierunków studiów na liście ministerialnej, skoro na liście tej znajduje się kierunek Informatyka i ekonometria, który jest związany z dwiema dziedzinami nauki: z naukami matematycznymi i/lub technicznymi – informatyka, oraz z naukami ekonomicznymi – ekonometria.

W podejmowaniu decyzji o racjonalności międzykierunkowego/makrokierunkowego programu kształcenia nie można zatem opierać się na sztywnych, formalistycznie traktowanych zasadach, nie da się bowiem określić takich zasad, które miałyby zastosowanie do wszystkich możliwych przypadków. W pewnych przypadkach opieranie kształcenia na odległych nawet dyscyplinach naukowych może być uzasadnione realną potrzebą interdyscyplinarności, w innych natomiast, może mieć znamiona zabiegu o charakterze de facto marketingowym, do czego szczególnie często bywa wykorzystywana informatyka. Uważamy, że jednym z przejawów nadużycia o tym charakterze jest wprowadzenie kształcenia pod nazwą *informatyka stosowana*. Na jakich bowiem dyscyplinach naukowych, poza informatyką, ma się takie kształcenie opierać? Na jakichkolwiek? Oczywiście wprowadzanie nauczania informatyki w ramach większości obszarów kształcenia jest dzisiaj niezbędne, ale powinno to być nauczanie zastosowań informatyki w konkretnym obszarze i mieć nazwę z tym obszarem związaną. Nauczanie to może być podobne dla wielu takich obszarów, ale to nie znaczy, że sensownym jest mówić o jego niezależności od obszaru kształcenia, a to właśnie sugeruje nazwa *informatyka stosowana*.

W dalszej części tekstu racjonalność kształceniu makrokierunkowego w obszarze NTI analizowana jest z punktu widzenia merytorycznego powiązania jego kierunków składowych oraz wynikającego z tego pożądanego charakteru i zawartości programów nauczania. Termin makrokierunek, który – jak już wspomniano – znalazł się w projekcie nowego prawa o szkolnictwie wyższym, został wprowadzony w połowie lat 90. przez RGSzW i jest używany w odniesieniu do kształcenia w zakresie NTI w Politechnice Warszawskiej.

Wzajemne przenikanie się treści nauczania z kierunków składowych NTI oczywiście następuje także bez formalnego wprowadzenia kształcenia na makrokierunku, ale w takim przypadku ma ono z reguły charakter wzbogacania programów nauczania poszczególnych kierunków, a nie konsekwentnego ich integrowania w ramach spójnej całości. Absolwenci konsekwentnie ukształtowanego makrokierunku powinni otrzymywać dyplom ukończenia studiów na tymże makrokierunku (a nie dyplom ukończenia studiów na którymś z jego kierunków składowych), z podaniem węższego obszaru wiedzy i umiejętności, w którym absolwent się specjalizował na starszych latach studiów i w fazie dyplomowania. W ramach NTI można tworzyć specjalności o nazwach pokrywających się z kierunkami

składowymi, na przykład, informatyka, ale także węższe specjalności/profile, takie jak oprogramowanie systemów informacyjnych czy systemy sterowania, a wreszcie – co szczególnie ważne – specjalności/profile o silnie podkreślonym charakterze interdyscyplinarnym, takie jak teleinformatyka (elementy telekomunikacji i informatyki) czy systemy multimedialne (elementy elektroniki, informatyki i telekomunikacji). Kształcenie na makrokierunku zapewnia zatem znaczną elastyczność profilowania specjalistycznych treści nauczania, a więc umożliwia szybkie reagowanie na zmiany zachodzące w związku z postępem technologicznym i na rynku pracy. Istotne jest także to, że zestaw specjalności/uprofilowań oferowanych w ramach makrokierunku może być skutecznie dopasowywany do specjalizacji i bieżących zainteresowań naukowych kadry danej jednostki akademickiej. W konsekwencji, jednostki akademickie mogą się istotnie różnić rodzajem i zakresem oferowanych specjalności/uprofilowań w ramach kształcenia na makrokierunku NTI, co sprzyja ich różnorodności, będącej ważnym atrybutem „zdrówego” szkolnictwa akademickiego.

Warunkiem sine qua non tego, by taka elastyczność specjalizowania/profilowania była osiągalna, jest to, by programy nauczania w pierwszej fazie studiów na makrokierunku zawierały, oprócz wiedzy ogólnej, także podstawy wiedzy specjalistycznej wszystkich kierunków składowych. Inaczej mówiąc, wszyscy studenci, niezależnie od tego, jaki profil specjalizacyjny obiorą w dalszych latach studiów, powinni nabyć wiedzę z zakresu podstaw wszystkich kierunków składowych makrokierunku NTI, tj. z zakresu automatyki, elektroniki, informatyki i telekomunikacji. To wymaganie stanowi główną trudność w konstruowaniu programów nauczania na pierwszych latach studiów I stopnia. Z oczywistych względów program taki nie może być „sumą” tradycyjnych programów nauczania podstaw specjalizacyjnych poszczególnych kierunków składowych. Konieczne jest więc opracowanie nowych programów nauczania. Nie jest to zadanie łatwe, gdyż nie wchodzi w grę prosta adaptacja programów przedmiotów stworzonych dla jednostopniowych i „jednokierunkowych” studiów magisterskich. Nowe programy nauczania tworzone w różnych środowiskach akademickich, będą zapewne się od siebie różnić i ewoluować wraz rozwojem i ewolucją wzajemnych powiązań dziedzin składowych NTI. W tym kontekście szczególnie ważny jest sposób, w jaki rozumie się istotę powiązań – konwergencji i przenikania się - dziedzin składowych NTI.

* * *

Biorąc pod uwagę przedstawione ogólne uwarunkowania rozwoju szkolnictwa wyższego związane z Procesem Bolońskim, a także bardziej szczegółowe kwestie dotyczące kształcenia w obszarze technik i technologii informacyjnych, można bez wątplenia stwierdzić, że przed polskimi uczelniami, mającymi ambicję zająć dobre miejsce w kształtującej się Europejskiej Przestrzeni Edukacyjnej, stoją poważne wyzwania organizacyjne i merytoryczne. Wyzwanie te muszą być podjęte niezwłocznie, w przeciwnym przypadku polskie uczelnie zostaną rychło zmarginalizowane. Podejmując te wyzwania warto uświadomić sobie nasze silne strony. Nie jest ich, niestety, zbyt wiele, tym bardziej powinniśmy je skutecznie wykorzystywać. Sądzymy, że najsilniejszą stroną naszego szkolnictwa wyższego, a w szczególności politechnicznego, jest to, że zachowało ono (jeszcze!) zdolność do realizowania programów nauczania opartych na solidnym wykształceniu podstawowym. To z tego powodu absolwenci naszych uczelni są doceniani na europejskim i amerykańskim rynku pracy. Nie zmarnujmy tego!

MATERIAŁY ŹRÓDŁOWE

- [Kras03] A. Kraśniewski, "Bolonia, Praga, Berlin, ... dokąd zmierza europejskie szkolnictwo wyższe", *Miesięcznik Politechniki Warszawskiej*, nr 12/2003 (wkładka).
- [Lubacz04] J. Lubacz, „Obszary, programy i standardy kształcenia – czy wiemy czego chcemy?”, *Forum Akademickie*, kwiecień 2004.
- [SEFI03] *Communication of CESEAR and SEFI on the Bologna Declaration*, Feb. 2003; www.ntb.ch/SEFI.
- [TrendsIII] *Trends in Learning Structures in Higher Education (III) – Progress toward the European Higher Education Area*, by S. Reichert and Ch. Tauch, July 2003; www.bologna-bergen2005.no.
- [Reding03] V. Reding, *Making European Higher Education a Worldwide Reference*, keynote address, EUA Convention of European higher education institutions, Graz, 29 May 2003, <http://eua.uni-graz.at>.
- [Tuning03] J. Gonzales, R. Wagenaar (eds.), *Tuning Educational Structures in Europe – Final Report, Phase I*, 2003; także http://europa.eu.int/comm/dgs/education_culture/Tuning.
- [CEC02] Education and training in Europe: The work programme on the future objectives of education and training in Europe, European Commission, 2002.
- [CEC03] The role of universities in the Europe of knowledge – Communication from the Commission, Commission of the European Communities, 5 Feb. 2003.
- [Ustawa04] Ustawa Prawo o szkolnictwie wyższym, projekt z dn. 22 stycznia 2004 r., <http://www.frp.org.pl>.

**EVOLUTION OF HIGHER EDUCATION IN EUROPE AND ITS IMPACT
ON EDUCATION IN THE AREA OF INFORMATION TECHNOLOGY****Abstract**

Current trends and future directions in the process of harmonisation of the systems of higher education in the European countries, referred to as the Bologna Process, are presented. In this context, several desirable characteristics of the system of engineering education in Poland, with particular emphasis on the education in the area of information technology, are formulated.

Wojciech Szpankowski

Department of Computer Science, Purdue University, W. Lafayette, IN 47907
<http://www.cs.purdue.edu/people/spa>

UBIQUITOUS PATTERN MATCHING AND ITS APPLICATIONS

Abstract

Repeated patterns and related phenomena in words are known to play a central role in many facets of computer science with applications in multimedia compression, information security, and computational biology. One of the most fundamental questions arising in such studies is the frequency of pattern occurrences in a string known as the text. Pattern matching comes in many flavors: In the *string matching* problem, for a given string (viewed as a consecutive sequence of symbols) one counts the number of pattern (string) occurrences in the text. In *subsequence pattern matching*, also known as the *hidden word problem*, we search for a given subsequence rather than a string. Finally, in *self-repetitive pattern matching* we aim to determine when a prefix of the text will appear again. We apply probabilistic and analytic tools of combinatorics and analysis of algorithms to discover general laws of pattern occurrences. An immediate consequence of our results is the possibility to set thresholds at which the appearance of a pattern in a given text starts being 'meaningful'. In this talk, we first demonstrate the application of string matching methodology to biological sequence analysis; in particular, to the problem of finding weak signals and avoiding artifacts. We then use the approach for hidden words to construct a reliable threshold for intrusion detection in detecting anomalies. Finally, we present a video compression scheme based on two-dimensional self-repetitive pattern matching (i.e., a lossy extension of the Lempel-Ziv scheme). We conclude this talk with a demo illustrating a real-time multimedia decoder based on mobile code that has potential application for wireless communication.

Henryk Krawczyk, Józef Woźniak

**Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska**

SYNERGIA I KONERGENCJA PODSTAWĄ ROZWOJU NOWOCZESNYCH TECHNOLOGII INFORMACYJNYCH

Streszczenie

Elektronika, telekomunikacja i informatyka są nowoczesnymi dyscyplinami nauki i techniki, które rozwijając się, w znacznej mierze niezależnie, wpływają na siebie wzajemnie, wprowadzając nowe, dodatkowe wartości (synergia) oraz kreując wspólną wizję świata cyfrowego (konwergencja i globalizacja). W pracy zaprezentowano zarówno stan rozwoju tych dyscyplin, jak i przyszłościową ich wizję. Zwrócono uwagę na nowe technologie elektroniczne, budowę wysokowydajnych i inteligentnych sieci komputerowych (sieci konwergentnych, w tym Internetu, szybkich sieci LAN, sieci wirtualnych, sieci łączności bezprzewodowej) oraz implementacje nowych modeli i architektur przetwarzania. Postawiono tezę, że synergia i konwergencja technologii informacyjnych gwarantuje szeroki dostęp do różnorodnych nowoczesnych usług teleinformatycznych, stanowiąc podstawę rozwoju społeczeństwa informacyjnego.

1. WSTĘP

Przełom nowego wieku wiąże się nierozdzielnie z powstawaniem społeczeństwa informacyjnego i z rozwojem technologii informacyjnych IT (*Information Technologies*), bądź też informacyjno-telekomunikacyjnych ICT (*Information and Communication Technologies*) [4] – ściśle związanych z elektroniką, telekomunikacją i informatyką. We wszystkich tych dyscyplinach obserwuje się, szczególnie w ostatnim 10-leciu, ogromny postęp w pracach badawczych i wdrożeniowych. W tabeli 1 zaprezentowano podstawowe atrybuty tego postępu, pozwalające ocenić dynamikę zmian zachodzących w obrębie technik IT. Poprawa szeregu występujących w tej tabeli parametrów, w tym wzrost gęstości upakowania układów elektronicznych, zwiększenie częstotliwości pracy układów (a w szczególności mikroprocesorów) oraz wzrost dostępnej wielkości modułów pamięci jest ogromną zasługą mikroelektroniki. Opracowanie nowych modeli przetwarzania czy zarządzania (efektywnych systemów operacyjnych, a także kreowania usług informacyjnych, bądź technik programowania (nowych języków i kompilatorów) jest wynikiem postępu w informatyce. Z kolei pojawienie się cyfrowych sieci komunikacyjnych, oraz wzrost szybkości transmisji wiadomości (nowe media i efektywne protokoły transportowe) jest niekwestionowaną zasługą telekomunikacji.

Dynamika rozwoju technologii informacyjnych w latach 1970 – 2005

Tabela 1.

Rok	Gęstość upakowania	Częst. zegara	Wielkość pamięci	Elementy Architektury model/program	Kompilatory	Systemy operacyjne	Uszczelnienie -- szybkość transmisji
1970	<1000	1MHz	1 KB	Sekwencyjny mikroprogramowanie słowo 8b	C, Pascal, analiza przepływu danych	Unix, podział czasu	ARPA LAN – 1Mb/s
1980	<10000	10MHz	<1MB	Współbieżny, potokowy mikroprogramowanie	GNU wektoryzacja środowiska programowania	RPC – remote procedure call DSF – distributed file systems	Internet TCP/IP Ethernet/Token Ring - 10/16Mb/s
1985	<50000	<50MHz	<32MB	CISC, RISC pamięć notatnikowa procesory personalne stacje robocze słowo 16b	ML, Te1/Tk analiza interproceduralna	Mikrojądra XWindows NFS – serwery nazw i plików danych	Protokoły zarządzania siecią FDDI – 100Mb/s
1990	<1 milion	100MHz	100MB	Równoległy superkomputery słowo 32b	Języki równoległe C++ HPF/ Fortran PVM, MPI	Obiekty rozproszone CORBA	ATM, HTTP Mbone MINE 150/622 Mb/s
1995	<10 milionów	500MHz	1GB	Rozproszony komputer sieciowy metakomputery słowo 64b	Java/JVM kod przenośny niezawodna kompilacja reuse	Kod przenośny, bezpieczeństwo WWW Agenty	Web, IPv6 Internet 2 sieci aktywne Gigabit Ethernet
2000	100 milionów	1GHz	100GB	Zespołowy wielowątkowość procesory wieloskalarne (klasy)	Programowanie WWW, kompilacje just in time	Windows, Serwery aplikacji	Internet nowej generacji, 1 bilion węzłów (WDM, 10Gb/s)
2005	1 miliard	10GHz	1TB	Układy wielomikroprocesorowe wielokomputerowe, komputery kwantowe i molekularne (gridy)	Interpretacje, kompozycje komponentów	Standardizowane usługi WWW, wirtualna rzeczywistość	Konwergencja systemów ruchomych i stacjonarnych

Można oczywiście dyskutować, które z wymienionych, jak też nie wymienionych elementów, mają szczególnie istotny wpływ na dynamikę rozwoju technologii informacyjnych. Nad wyraz trudno byłoby ustalić jednoznaczne kryteria ich ważności. Nie to jest też najistotniejsze z punktu widzenia niniejszego opracowania. Najważniejsze jest spostrzeżenie, że motorem rozwoju wszystkich wymienionych wcześniej dyscyplin są dwie podstawowe wartości, a mianowicie konwergencja i synergia! Zatem całość analizowanej problematyki rozpatrywana będzie w kontekście tych wartości.

Konwergencja wiąże się z tendencją do tworzenia systemów o uniwersalnych cechach, a także o podobnej budowie i własnościach funkcjonalnych. W przypadku technologii informacyjnych uwidacznia się ona w oferowaniu przez współczesne systemy różnorodnych usług, aplikacji i systemów, charakterystycznych dotąd dla odrębnych rozwiązań.

Synergia oznacza z kolei współdziałanie różnych elementów systemu prowadzące do wzmocnienia efektywności i skuteczności funkcjonowania całego systemu, jak również pojawienia się nowych, dotychczas nieznanych własności czy możliwości. Dobrym przykładem synergii, w przypadku IT, jest nowe środowisko przetwarzania WWW, oferujące znacznie większe możliwości – w porównaniu z możliwościami tradycyjnych sieci komputerowych. Nie ma jak dotąd ścisłych matematycznych modeli opisujących konwergencję i synergję. Nie można więc analizować wpływu tych wartości w sposób zbyt formalny.

Skupimy się więc na nieformalnych podejściach, które pozwalają lepiej zrozumieć podstawowe warunki rozwoju IT, jak też określić możliwe kierunki ich rozwoju w przyszłości. Łatwo zauważyć, że nowe technologie, związane z Internetem, protokołami IP, telefonią komórkową, systemami operacyjnymi, zaawansowanymi aplikacjami rozwijają się burzliwie dzięki sukcesom mikroelektroniki, optoelektroniki czy też techniki mikrofalowej. Synergia szerokiego wachlarza dyscyplin IT, wzmacnia również skuteczność i efektywność działań podejmowanych przez projektantów i producentów. W jej wyniku obserwujemy rozwój nowych systemów, usług sieciowych i aplikacji. Jednocześnie znacznemu obniżeniu ulegają koszty budowy systemów teleinformatycznych i dostarczania nowych usług. Rośnie też dostępność tych usług i powszechność ich akceptacji. Różne instytucje zainteresowane są korzyściami płynącymi z inwestowania w „narzędzia” informatyczne i aplikacje. Przyspiesza to z kolei procesy „odnawiania się” technologii IT. Wielkie, rewolucyjne zmiany w technologiach informacyjno-telekomunikacyjnych, obserwowane w ostatnich latach, związane są właśnie z konwergencją różnego typu systemów. Motorem tych zmian, w znacznym stopniu są również procesy „biznesowe” [3].

Celem niniejszej pracy jest więc próba naszkicowania aktualnego stanu informatyki, telekomunikacji i elektroniki – dyscyplin decydujących o rozwoju wielu gałęzi techniki i biznesu, a także wskazanie roli jaką odgrywać one będą w kształtowaniu się społeczeństwa informacyjnego. Wykorzystując nowe technologie elektroniczne, zaawansowane systemy telekomunikacyjne oraz wydajne aplikacje informatyczne można budować złożone, zintegrowane systemy użytkowe o strukturach wielowarstwowych. Definiując funkcje poszczególnych warstw, implementujące je mechanizmy, a także interfejsy międzywarstwowe (modułarność i standaryzacja) radykalnie upraszcza się procesy: projektowania, wytwarzania i zarządzania tymi systemami. Umożliwia to również wykorzystanie różnego rodzaju narzędzi wspomagających te procesy – co prowadzi zarówno do poprawy elastyczności funkcjonowania systemów teleinformatycznych, jak też automatyzacji ich projektowania.

2. UKŁADY I SYSTEMY ELEKTRONICZNE

Jak już wspomniano, elektronika dostarcza układów do budowy różnego typu systemów cyfrowych, w tym sieci komputerowych. Z dotychczasowych obserwacji wynika, że mniej więcej co trzy lata pojawiają się nowe technologie informacyjne. Powodują one każdorazowo zwiększenie szybkości pracy układów cyfrowych, podniesienie niezawodności ich działania i zmniejszenie kosztów wytwarzania. I tak na przykład, w porównaniu do roku 1970, niezawodność układów w roku 1999 wzrosła aż 10^4 razy, a koszt wytwarzania zmniejszył się dokładnie w tym samym stopniu. Obecnie w klasycznej elektronice (układów i podzespołów), która przyjęła nazwę mikroelektroniki (z uwagi na rozmiary produkowanych układów), wyróżnia się cztery podstawowe rodzaje wyrobów [2]:

- standardowe układy scalone,
- specjalizowane układy scalone,
- programowalne układy scalone,
- mikromechanizmy i mikrosystemy krzemowe.

Standardowe układy, zwane potocznie katalogowymi, są produkowane w wielkich seriach i przeznaczone do rynkowej sprzedaży w dowolnych ilościach i dowolnym odbiorcom. Należą do nich między innymi mikroprocesory i pamięci. Ich produkcja wymaga znacznych nakładów inwestycyjnych. Dzięki istniejącej konkurencji parametry tych układów są stale ulepszane (patrz tabela 1 dotycząca układów mikroprocesorowych). Zauważa się ogólne tendencje, które stwierdzają, że co trzy lata:

- powierzchnia najmniejszego możliwego do wykonania układu zwiększa się 1,4 razy,
- maksymalna liczba elementów (tranzystorów) w układzie zwiększa się 6-krotnie,
- maksymalna szybkość działania zwiększa się 3 razy.

Przewiduje się, że w roku 2010, przy produkcji układów krzemowych, osiągnięte zostaną nieprzekraczalne bariery technologiczne i materiałowe – wynikające z praw fizyki. Dalszy postęp będzie możliwy, jeśli zostaną znalezione i wykorzystane zupełnie nowe materiały. Coraz więcej mówi się o układach molekularnych, kwantowych, optycznych czy biologicznych (genetycznych). Wówczas będziemy mieli do czynienia nie z mikroelektroniką, a z tzw. nanosystemami, które charakteryzują się całkiem innymi zasadami budowy i przetwarzania informacji [5]!

Specjalizowane układy scalone (ASIC – *Application Specific Integrated Circuits*) są projektowane i wytwarzane dla potrzeb konkretnego użytkownika. Nie występują więc w standardowych katalogach, a ich produkcja dotyczy niewielkich ilości. Z uwagi na mniej złożoną technologię wytwarzania koszt przygotowania produkcji jest mniejszy w porównaniu do takich kosztów wytwarzania układów standardowych. Obserwuje się, że ich udział w produkcji podzespołów elektronicznych jest obecnie stale rosnący i obecnie wynosi już ponad 50%.

Układy scalone programowalne, podobnie jak układy ASIC (często zresztą błędnie określane tą samą nazwą), są przeznaczone dla jednego odbiorcy, jako elementy do wytwarzanego przez niego sprzętu. Z punktu widzenia produkcji są to układy standardowe, z punktu widzenia użytkownika, który określa ich funkcję już po ich wyprodukowaniu, są układami specjalizowanymi. Dlatego też są one z jednej strony dość kosztowne, z drugiej zaś dzięki funkcjonowalnej nadmiarowości, łatwe do zaprogramowania.

Ostatnia, czwarta grupa wytwarzanych produktów elektronicznych jest dobrym przykładem wykorzystania zasady konwergencji. Mikromechanizmy i mikrosystemy krzemowe łączą cechy zarówno układów elektronicznych, jak i elementów mechanicznych, czy nawet zjawisk chemicznych. Przykładami mogą być czujniki różnych wielkości fizycznych (czujniki drgań, stężenia gazu), czy mechanizmy wykonawcze (takie jak mikrosilniki, mikropompy, czy mikrogrzejniki). Rośnie przy tym szybko liczba zarówno nowych rozwiązań, jak i możliwości ich zastosowań (np. jako czujniki przyspieszenia – wyzwalające poduszki powietrzne w samochodach, urządzenia przenośne w telekomunikacji, w tym telefony komórkowe). Szacuje się, że w roku 2006 liczba „popularnych komórek”, użytkowanych na całym świecie, przekroczy 1 miliard!

Rynek obecny powoli nasycy się układami standardowymi i coraz większą w nim rolę odgrywać będą układy specjalizowane i mikrosystemy, czy nanosystemy. Roczny przyrost produkcji urządzeń biurowych wynosi 8-10%, domowych 10-15%, zaś ruchomych 15-20%. W ostatnim dwudziestoleciu mikroelektronika stała się jednym z najważniejszych elementów współczesnej cywilizacji technicznej.

W Polsce rynek ten po przemianach polityczno-gospodarczych w latach 80-tych, wyraźnie się załamał. Potwierdza to upadek Centrum Mikroelektroniki (CEMI), zaś pewną nadzieją jest przetrwanie Instytutu Technologii Elektronicznej (ITE) i Instytutu Technologii Materiałów Elektronicznych (ITME), które starają się włączyć do współpracy międzynarodowej. Dobrym sygnałem w tym względzie są również inwestycje zagraniczne w województwie pomorskim. Chodzi tu przede wszystkim o inwestycję firmy Flextronics International (FI), która buduje swój zakład w specjalnej strefie ekonomicznej Tczew-Zarnowiec.

Postęp w rozwoju układów elektronicznych, głównie tych standardowych, miał ogromny wpływ na rozwój różnego typu systemów, w tym architektur systemów telekomunikacyjnych i systemów komputerowych (patrz tabela 1). Oprócz tradycyjnych modeli przetwarzania (przetwarzanie sekwencyjne, współbieżne) pojawiły się nowe modele (przetwarzanie równoległe i rozproszone), których celem jest zwiększenie zarówno wydajności, jak i wiarygodności działania, tj. przyspieszenia wykonywania operacji, jak i wzrost pewności poprawnego jej wykonania. Modele te są implementowane i wykorzystywane nie tylko w złożonych systemach wieloprocesorowych, czy wielokomputerowych, ale obecnie lokowane są również w pojedynczym komputerze. Przykładem tego jest choćby procesor Pentium, w którym występuje zarówno wielostrumieniowe przetwarzanie potokowe, jak i przetwarzanie równoległe, gdzie w tym samym czasie przesyłane są różne strumienie danych i wykonywane są na nich różne funkcje (instrukcje). Jest to możliwe dzięki ogromnemu upakowaniu tranzystorów w jednym układzie. Obecnie proponuje się już architektury systemów obejmujące 1 miliard tranzystorów, które staną się bazowymi komputerami w latach 2010.

Inną obserwowaną tendencją jest rozwój przetwarzania rozproszonego. Jest to przykład konwergencji i synergii pomiędzy architekturą komputerów a sieciami komputerowymi. Dzięki coraz to szybszej komunikacji istnieje możliwość rozproszenia obliczeń między różne węzły sieci. Takie podejście może zwiększyć wydajność systemu (poprzez np. równoległość obliczeń) bądź jego wiarygodność (poprzez replikację obliczeń). Zmienia się również sposób zarządzania wykonywaniem tego typu zadań. Systemy operacyjne pojedynczych węzłów pracujących autonomicznie muszą mieć możliwość koordynowania pracy systemu jako całości. Otwierają się więc nowe możliwości w sferze organizacji procesu zarządzania. Okazuje się, że coraz bardziej upodabniają się one do procedur zarządzania zespołami ludzkimi, zwłaszcza, gdy w funkcjonowanie systemu włączone są,

na różnych poziomach, decyzje ludzkie. Mówimy wówczas o tzw. przetwarzaniu zespołowym, czego przykładem może być wykonywanie ściśle określonych zadań związanych z załatwianiem określonej sprawy administracyjnej. Petent pozostaje w domu, a niezbędne informacje może uzyskać poprzez sieć oraz pośredni kontakt z odpowiednimi urzędnikami. W trakcie takiej sesji mogą być na bieżąco podejmowane decyzje (ludzkie) o dalszych krokach postępowania. Ten model przetwarzania dotyczący obliczeń zespołowych, integrujący różnych specjalistów dla załatwienia konkretnej sprawy, będzie w przyszłości szeroko rozwijany, ponieważ stanowi podstawę tworzenia tzw. usług społeczeństwa informacyjnego, gdzie informacja i decyzje są głównymi elementami jego funkcjonowania.

3. KONWERGENTNE SIECI TELEKOMUNIKACYJNE

Współczesne sieci telekomunikacyjne oferują integrację usług, umożliwiając obsługę różnych typów ruchu. W chwili obecnej przekaz danych, charakterystyczny dla przekazów komputerowych, zaczyna przewyższać swą objętością ruch rozmówny – typowy dla telefonii. Sprawia to, że nowe, rozwijane obecnie, technologie sieci przewodowych (a także bezprzewodowych) zakładają konwergencję mowy i danych (ogólnie: danych i informacji multimedialnych). Dotyczy to w zarówno systemów ATM, jak też nowych generacji sieci IP (czyli Internetu). Podstawową trudnością, którą należy rozwiązać w przypadku „sieci konwergentnych” jest spełnienie przez nie zróżnicowanych wymagań jakościowych „narzuconych” przez różne wymagania użytkowników i realizowanych przez nich aplikacji. W przypadku transferu głosu i obrazów podstawowymi atrybutami jakości będą: dopuszczalne opóźnienie i zmienność tego opóźnienia. Z kolei przy transferze danych krytycznym parametrem staje się prawdopodobieństwo utraty pakietu.

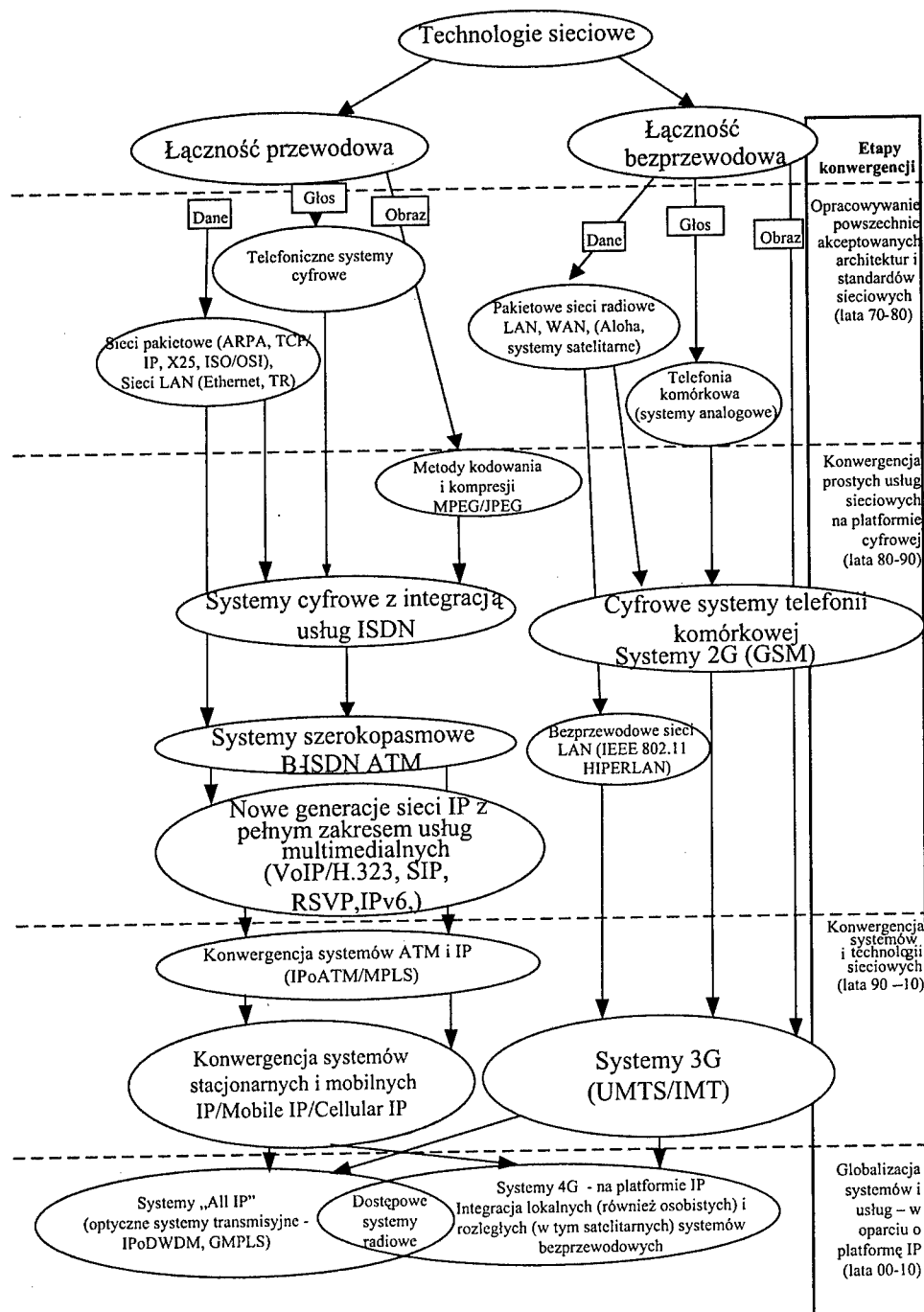
Sieci konwergentne oferować więc winny różnorodne usługi i zapewniać obsługę różnych typów ruchu. Przykładem takiej konwergencji jest świadczenie przez jedną sieć usług telemedycznych, teledoradztwa, zdalnego nauczania – teleedukacji oraz coraz powszechniejszych usług telebiznesu. Wiąże się to oczywiście z integrowaniem istniejących i przyszłych technologii sieciowych – stając się ogromnym wyzwaniem dla operatorów, usługodawców i projektantów. Sieci teleinformatyczne winny być bowiem „odporne i przeźroczyste na zmiany”, a zatem „podatne” na nowe technologie, a jednocześnie ciągle efektywne [1]. Wynikiem pracy organów standaryzujących Unii Europejskiej jest „Zielona Księga o Konwergencji Systemów Informacyjnych” (*Green Papers on the Convergence of Telecommunication Systems*). Dokument ten koncentruje się na nowoczesnych usługach cyfrowych i rozwiązaniach infrastruktury sieciowej. Procesy integracji usług i technologii sieciowych, obserwowane na przestrzeni ostatnich lat, pokazane są na rys. 1.

Wymagania dotyczące sieci teleinformatycznych, w tym i Internetu, ulegają ciągłym zmianom. W ostatnich latach coraz większe znaczenie przywiązuje się do gwarancji jakości (QoS) i niezawodności oferowanych usług. Podobnie jak i w innych rodzajach sieci telekomunikacyjnych, również i tutaj mówi się o konwergencji usług, czyli integracji obsługi w jednej sieci, różnych typów ruchu w szczególności: danych, głosu i obrazów. Zapewnienie wielousługowego charakteru sieci IP wymaga oczywiście zdefiniowania i wprowadzenia zróżnicowanych usług sieciowych, podobnie jak ma to miejsce np. w sieci ATM. Usługi te powinny różnić się pomiędzy sobą takimi parametrami jak: oferowana szybkość przekazu, szczególnie w warunkach przeciążenia sieci, gwarantowane opóźnienie przekazu i wartość zmienności tego opóźnienia (ang. *jitter*), typ przekazu (o stałej czy zmiennej szybkości) itd.

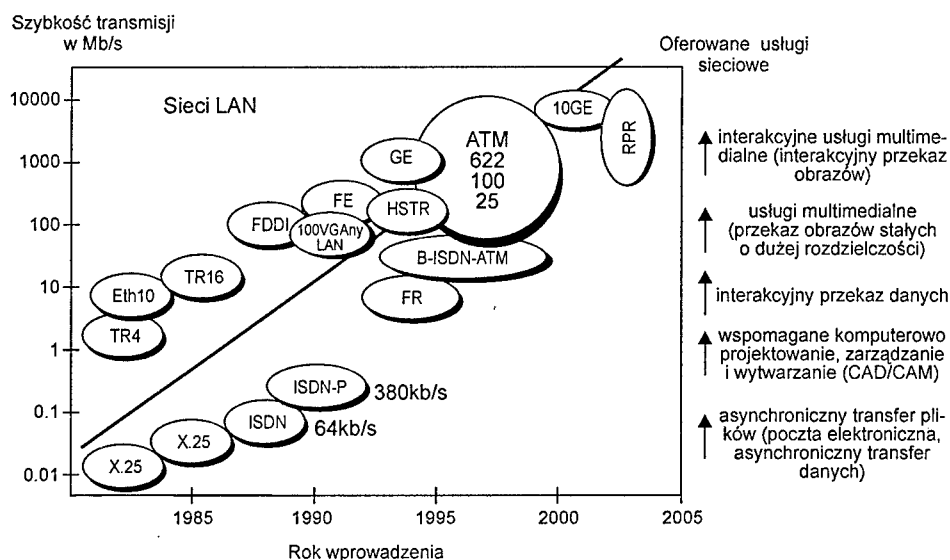
W sieciach wielousługowych wyróżniamy dwa podstawowe rodzaje ruchu, tj. strumieniowy i elastyczny. Ruch strumieniowy jest generowany przez aplikacje związane z przekazem obrazów ruchomych (np. wideo) czy też dźwięku (audio, Voice over IP). Ruch ten powinien być przesyłany przez sieć z małym (pomijalnym) opóźnieniem, przy jednoczesnym zapewnieniu niskich strat. Z kolei, ruch elastyczny dotyczy przesyłania dokumentów takich jak pliki, nieruchome obrazy itd. Dla tego rodzaju ruchu wymaga się zapewnienia poprawności przekazu, natomiast wymagania nakładane na czas przekazu nie są traktowane jako szczególnie krytyczne. Inaczej mówiąc aplikacje typu strumieniowego są to aplikacje wrażliwe na opóźnienia czasowe, podczas gdy aplikacje elastyczne charakteryzują się małą wrażliwością na wielkość i zmienność opóźnienia.

Obecna generacja sieci Internet, wykorzystująca głównie protokół IPv4, realizuje przekaz datagramów zgodnie z zasadą *best effort* – czyli „największego wysiłku” (określaną też często mianem zasady „największej staranności”), co właściwie nie zapewnia pożądanej jakości obsługi. Sieć ta nie realizuje też żadnej polityki przyjmowania czy odrzucania nowych połączeń, zaś sterowanie liczbą pakietów napływających do sieci jest scedowane na systemy końcowe. De facto, mechanizmy zaimplementowane w protokole TCP (takie jak mechanizm wolnego startu oraz unikania przeciążenia) mają na celu jedynie dostosowanie szybkości wysyłania pakietów do sieci, do aktualnie obserwowanych w niej warunków ruchowych. Z myślą o przekształcaniu sieci Internet w pełni wielousługową sieć IP zostały zaproponowane dwie nowe architektury sieci. Pierwsza z nich związana jest z realizacją tzw. Usług Zróżnicowanych – architektura DiffServ (ang. *Differentiated Services*), druga natomiast dotyczy tzw. Usług Zintegrowanych – IntServ (ang. *Integrated service*).

Należy jednakże podkreślić, że pierwsze prace nad sieciami konwergentnymi związane były z opracowaniem rozwiązań wąsko- i szerokopasmowych sieci ISDN (*Integrated Services Digital Network*) oraz sieci ATM (*Asynchronous Transfer Mode*). U podłoża procesów konwergencji leży bowiem integracja technologii cyfrowego przetwarzania i przekazu informacji, stanowiąca uniwersalne narzędzie obróbki sygnałów. Z kolei efektywne metody kompresji informacji, w tym standardowe algorytmy typu MPEG, czy też JPEG – stosowane przy przekazie obrazów stałych i ruchomych, pozwalają znacznie ograniczyć niezbędne przepustowości wykorzystywanych kanałów transmisyjnych. Sieci ATM oferują obsługę wielu klas ruchu, w tym CBR (*Constant Bit Rate*), VBR (*Variable Bit Rate*), ABR (*Asynchronous Transfer Mode*) i UBR (*Unspecified Bit Rate*) (zgodnie z klasyfikacją ATM Forum) gwarantując negocjowaną jakość obsługi. Klasy te odwzorowywane są na aplikacje związane z przekazem głosu, obrazów oraz (bardziej i mniej) pilnych danych. Sieci ATM pozwalają obecnie na realizację transmisji z szybkościami od 25 do 622 (i więcej) Mb/s. Z uwagi na ich ogromne możliwości prowadzono zaawansowane prace nad opracowaniem procedur współpracy sieci ATM i IP (Internetu). Prace te realizowane były przez zarówno IETF, jak też firmy zrzeszone w ATM Forum. Ich efektem są standardy IPoATM (*IP over ATM*), MPOA (*Multi-Protocol over ATM*), LANE (*LAN Emulation*), a ostatnio MPLS (*Multi-Protocol Label Switching*) i GMPLS (*Generalized MPLS* – z wykorzystaniem DWDM) – rozwiązania oparte na szybkiej komutacji etykiet i pozwalające na współpracę szeregu protokołów, w tym IP i IPX ze standardem ATM.



Rys. 1. Ilustracja procesu integracji usług i technologii sieciowych



Rys.2. Ewolucja technologii sieciowych

Ozn.: Eth10: Ethernet 10Mb/s, TR4(16): Token Ring 4Mb/s (16 Mb/s), FE: Fast Ethernet (100 Mb/s), GE: Gigabit Ethernet (1000 Mb/s), HSTR: High Speed Token Ring (100 Mb/s – 1000 Mb/s), FDDI: Fiber Distributed Data Interface (100 Mb/s).

Również w odniesieniu do popularnych i powszechnie stosowanych sieci LAN (*Local Area Networks*) mówić można o konwergencji usług. Współczesne sieci LAN – to bowiem coraz częściej LANY „multimedialne”, oferujące transfer danych, obrazów stałych i ruchomych oraz obsługę, przynajmniej częściowo, różnorodnych interakcyjnych aplikacji czasu rzeczywistego. Wszystko to jest wynikiem stosowania nowych typów procesorów, wykorzystania szerokopasmowych mediów transmisyjnych, użycia superszybkich przełączników (warstwy 2 i 3) i routerów, jak też opracowania efektywnych protokołów transportowych.

Dynamikę rozwoju zarówno sieci LAN, jak i sieci rozległych WAN (*Wide Area Networks*), z uwzględnieniem szybkości transmisji i obsługiwanych aplikacji, ukazuje rys. 2. Tendencje podobne do opisanych powyżej dotyczą też sieci łączności bezprzewodowej. Sieci te, rozwijające się dynamicznie w ostatnich 15-latach, są odpowiedzią producentów na coraz mocniej akcentowaną potrzebę obsługi użytkowników ruchomych. Postępy w mikroelektronice, w tym miniaturyzacja komponentów i urządzeń końcowych, pozwala na wytwarzanie małowymagowych i energooszczędnych terminali. Upowszechnianie się systemów i usług trzeciej generacji (3G) sprawi zapewne, że „osobisty” Internet i „osobiste” usługi multimedialne będą siłą napędową technologii IT w najbliższych latach. Obserwuje się, że dynamika rozwoju sektora usług osobistych (czytaj oferowanych przez systemy bezprzewodowe) jest zdecydowanie większa od dynamiki zmian w obszarze „połączeń stałych”. Z faktem tym wiąże się też istotne zmiany w technikach wielodostępu i modulacji. Po okresie dominowania techniki TDMA (*Time Division Multiple Access*) ogromne nadzieje wiąże się z wielodostępem kodowym CDMA (*Code Division Multiple Access*). W szybkich sieciach WLAN i WATM (*Wireless ATM*) wykorzystuje się też techniki rozpraszania widma DS SS (*Direct Sequence Spread Spectrum*) i FH SS (*Frequency*

Hopping SS), a także modulacje wielotonowe typu OFDM (*Orthogonal Frequency Division Multiplexing*). Przed sektorem łączności bezprzewodowej stoją ogromne wyzwania dotyczące koniecznych działań standaryzacyjnych i rozwoju nowych technologii. Kluczowy będzie tu zapewne wybór platformy telekomunikacyjnej. Wszystko wskazuje na to, że będzie to platforma IP. Nadzieje wiąże z nią nie tylko „Świat Internetu” ale także „Świat Telekomunikacji”.

Równie intensywnie prowadzone są też prace nad innymi systemami łączności bezprzewodowej, w tym sieciami WLAN (*Wireless LAN*). Prace te, prowadzone zarówno przez ETSI (*European Telecommunications Standards Institute*), jak też IEEE i inne krajowe, bądź międzynarodowe gremia normalizacyjne, zmierzają do opracowania kilku standardowych rozwiązań sieciowych o zróżnicowanych właściwościach. Część z tych rozwiązań opracowywanych jest z myślą o bezprzewodowej współpracy z przewodowymi systemami ATM. We wszystkich pracach standaryzacyjnych zauważa się wykorzystywanie coraz wyższych pasm częstotliwości (5 i 17 GHz – w standardzie HIPERLAN) oraz realizację transmisji z szybkościami sięgającymi 54 Mb/s (standardy: IEEE 802.11a i HIPERLAN 2), a nawet 108 Mb/s (aktualnie w fazie eksperymentów – są też prace nad rozwiązaniami oferującymi transfer z szybkością rzędu 500 Mb/s!).

W pierwszej fazie swego rozwoju sieci WLAN były postrzegane głównie jako alternatywa dla rozwiązań przewodowych w miejscach, w których instalacja tradycyjnych sieci LAN była niemożliwa (np. w obiektach zabytkowych) lub nieopłacalna ekonomicznie (np. tymczasowe sieci na potrzeby konferencyjne). Z biegiem czasu sieci WLAN zaczęły zdobywać szerszą akceptację i obecnie konkurują z rozwiązaniami tradycyjnymi w miejscach dotychczas zarezerwowanych dla sieci przewodowych, takich jak biura, zakłady pracy czy publiczne punkty dostępu do Internetu.

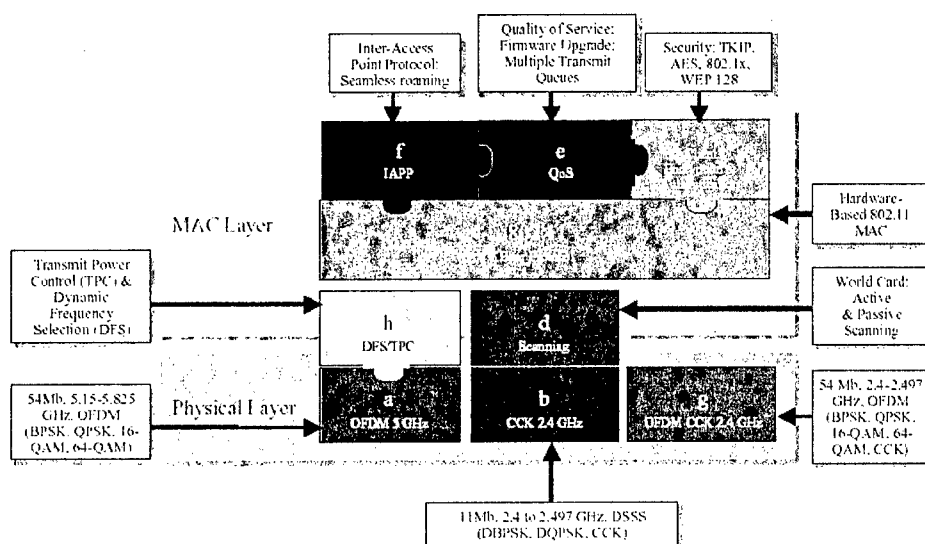
Istotnym punktem zwrotnym w rozwoju technologii WLAN było wprowadzenie systemów telekomunikacji ruchomej drugiej generacji (GSM). Sukces komercyjny systemów 2G wywarł ogromny wpływ na upowszechnienie komunikacji bezprzewodowej. Wraz z rozwojem systemów trzeciej generacji (UMTS) nastąpi, niewątpliwie upowszechnienie usług szerokopasmowych, w tym przekazów multimedialnych. Obecnie sieci WLAN postrzega się jako jeden z elementów systemów łączności bezprzewodowej czwartej generacji (4G).

Technologie sieci WLAN obejmujące szeroką gamę propozycji (od sieci osobistych typu Bluetooth, po klasyczne sieci WLAN i sieci dostępowe) mają liczną rzeszę zwolenników. Realizowane obecnie prace projektowe jak i standaryzacyjne dotyczą: technik antenowych, zaawansowanych szerokopasmowych technik nadawczych, metodą dostępu do medium i mechanizmów QoS, a także zagadnień bezpieczeństwa pracy pomiędzy punktami dostępu. Najpopularniejszym rozwiązaniem, powszechnie stosowanym w praktyce jest standard IEEE 802.11, ze swymi licznymi uzupełnieniami. Jednym z istotnych rozwiązań tego standardu jest propozycja 802.11e, rozbudowująca zasady dostępu do medium o mechanizmy QoS – pozwalające na różnicowanie obsługi różnych klas ruchu.

Istniejące rozwiązania WLAN charakteryzują się niską efektywnością obsługi QoS. Z tego powodu niezbędne jest opracowanie nowych algorytmów dostępu do kanału radiowego, które będą efektywnie zarządzały pasmem radiowym, zgodnie z wymaganiami aplikacji czasu rzeczywistego. W szczególności dotyczy to standardu IEEE 802.11. Zdając sobie z tego sprawę, organizacja IEEE powołała w roku 2000 grupę roboczą 802.11e, której celem było rozszerzenie warstwy MAC standardu IEEE 802.11 o mechanizmy QoS. Przegląd istotnych zagadnień związanych z pracą sieci WLAN, w tym w szczególności, standardu 802.11 (według stanu z 2003 r.) ilustruje rys. 3. Ciekawe propozycje zgłaszane są

też przez ETSI. Instytucja ta pracuje nad grupą rozwiązań typu HIPERLAN. Najbardziej obiecujące są prace nad nową generacją sieci WLAN, znaną pod nazwą HIPERLAN/2, i kompatybilną z rozwiązaniami ATM.

Inny kierunek prac wiąże się z systemami satelitarnymi nisko-, średnio- i wysoko-orbitowymi. Wspomagać one będą zarówno typowe przekazy multimedialne między użytkownikami stacjonarnymi (np. sieci VSAT – *Very Small Aperture Terminals*), jak też, a może przede wszystkim, między użytkownikami ruchomymi. „Sieci” satelitów nisko-orbitowych (por. np. Iridium) gwarantują małe opóźnienia czasowe, globalną lokalizację użytkowników i stałe śledzenie ich przemieszczania się, z możliwością szybkiego przełączania połączeń. Z uwagi na ogromny postęp w technologii światłowodów i opracowanie techniki wielodostępu z podziałem długości fali (WDM i DWDM) dynamika rozwoju systemów satelitarnych uległa jednakże pewnemu przyhamowaniu.



Rys. 3. Rodzina standardów serii IEEE 802.11

Pojęciem, które nierozdzielnie wiąże się z nowoczesnymi technologiami jest szeroko pojęta mobilność. Popularność telefonii bezprzewodowej wymusza nowe standardy we wszystkich dziedzinach techniki związanych z zapewnieniem łączności przemieszczającemu się użytkownikowi globalnej sieci informacyjnej. Rozpowszechnienie technologii bezprzewodowych w komunikacji wpłynęło również na tworzenie nowych standardów w dziedzinie sieci komputerowych. Nowe interfejsy sieciowe zapewniające wydajne połączenia przy zastosowaniu medium radiowego stały się pierwszym krokiem w kierunku zapewnienia użytkownikowi sieci komputerowej pełnej mobilności. Kolejnym jest niewątpliwie stworzenie nowych technologii umożliwiających bezprzewodową komunikację pomiędzy przemieszczającymi się użytkownikami. Następstwem tych przemian jest konieczność opracowania rozwiązań umożliwiających „przezroczyste” włączenie ruchomych użytkowników bezprzewodowych sieci komputerowych do sieci globalnej. Zadanie to wymaga opracowania zarówno protokołów przekazywania danych do i od użytkownika ruchomego jak również standardów związanych z zarządzaniem strukturą komputerowej sieci bezprzewodowej. Sieć Internet stanowiąca dominujący standard w sieciach kompu-

terowych opiera swoje działanie na protokole warstwy sieciowej IP – Internet Protocol. Protokół ten wymusza realizację określonych procedur i zastosowanie określonych norm związanych między innymi z adresowaniem węzłów sieci.

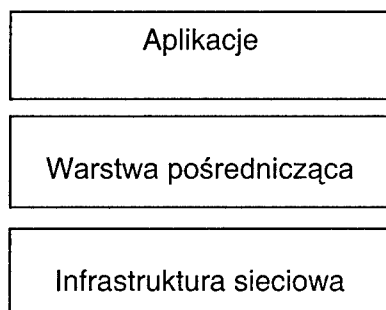
Konieczność zdefiniowania protokołu wspierania mobilności w sieci Internet doprowadziła do opracowania przez organizację IETF (*Internet Engineering Task Force*) standardu Mobile IP. Protokół, ten wykorzystujący koncepcję zastosowania agentów mobilności, definiuje wszystkie niezbędne elementy związane z zapewnieniem obsługi przemieszczającego się użytkownika sieci Internet. Procedury realizowane przez MIP umożliwiają określenie lokalizacji stacji ruchomej, przesyłanie pakietów zarówno do jak i od stacji oraz zapewniają bezpieczeństwo przesyłanych danych. Protokół IETF Mobile IP jest stale rozwijany. Propozycja rozszerzenia standardu IETF Mobile IP with Routing Optimization wprowadza dodatkowe procedury usprawniające przesyłanie pakietów. W oparciu o IETF Mobile IP tworzone są również nowe koncepcje. Przykładem może być Scalable Mobile Host IP Protocol.

Protokół IETF Mobile IP stanowiący podstawowy standard dla realizacji sieci bezprzewodowych posiada wady wynikające z przyjętych w trakcie tworzenia założeń. W celu usprawnienia procesu przesyłania danych w sieci oraz zmniejszenia obciążenia protokółarnego w sieci szkieletowej zaproponowane zostały między innymi protokoły: Cellular IP i HAWAII (*Handoff Aware Wireless Internet Infrastructure*). Sieci pracujące w oparciu o te protokoły pełnią funkcję bezprzewodowych sieci dostępowych dla przemieszczających się w jej obrębie stacji. Sposób działania tych protokołów znacznie odbiega od rozwiązań stosowanych w IETF Mobile IP i jest bliższy działaniu sieci bezprzewodowej telefonii komórkowej.

4. SYSTEMY, APLIKACJE, USŁUGI INFORMATYCZNE

Specjaliści z dziedziny technik informacyjnych wytwarzają złożone produkty, którymi są różnego rodzaju aplikacje, usługi czy złożone systemy informatyczne. Terminy: aplikacje, usługi, systemy są często różnie interpretowane, bądź nawet stosowane zamiennie. W celu uniknięcia niejednoznaczności terminologicznych przyjmijmy następujące definicje:

- System informatyczny jest zbiorem ściśle zdefiniowanych i powiązanych z sobą zadań użytkowych i określonych zasobów systemowych. System udostępnia zadaniom użytkowym swoje zasoby i usługi, umożliwia komunikację, nadzoruje ich wykonanie oraz podtrzymuje współpracę z użytkownikami. Przykład: informatyczny system handlu elektronicznego, czy komputerowy system bankowy.
- Aplikacja użytkowa jest to zbiór dobrze określonych modułów programowych (komponentów) powiązanych określonym celem użytkowym, działających w systemie komputerowym i korzystających z jego dostępnych zasobów i usług, na ogół definiowanych jako API (*Application Program Interface*). Przykład: realizacja konkretnego rodzaju zakupów w Internecie, zapłata kartą kredytową, czy zarządzanie bazą produktów.
- Usługi systemowe stanowią zbiór ściśle określonych czynności zaimplementowanych w systemie, bądź jako biblioteki funkcji systemowych, bądź jako zestaw podstawowych mechanizmów oferowanych przez różne komponenty systemu. Przykład: mechanizmy cookies do synchronizacji operacji zakupów, czy wyświetlanie pocztu w standardzie HTML.



Rys. 4. Podstawowe warstwy sieciowego systemu informatycznego

Z powyższych definicji wynika, że aplikacje na poziomie niższym mogą stanowić usługi dla aplikacji na niej nadbudowanej. Z kolei system informacyjny może być postrzegany jako zestaw wykonywanych aplikacji, bądź zbiór dostępnych usług albo jako zbiór jednocześnie dostępnych aplikacji i usług. Tak więc użyteczność systemu informatycznego może być oceniana poprzez możliwość wykonania żądanej usługi, bądź konkretnej aplikacji użytkowej. Do opisu systemu informatycznego, czy sieci komputerowej wykorzystuje się na ogół model warstwowy. Liczba warstw w takim modelu zależy od jego złożoności funkcjonalnej. Na ogół warstwę najwyższą stanowi aplikacja – wyznaczona poprzez konkretną platformę programistyczną, zaś warstwę najniższą warstwa fizyczna – zaimplementowana w dostępnej technologii elektronicznej (Rys. 4 przedstawia trójwarstwowy model systemu informatycznego). Modele warstwowe wykorzystuje się również przy opisie różnego typu aplikacji czy usług. Istnieją dwie różne szeroko rozpowszechnione klasy dostawców usług:

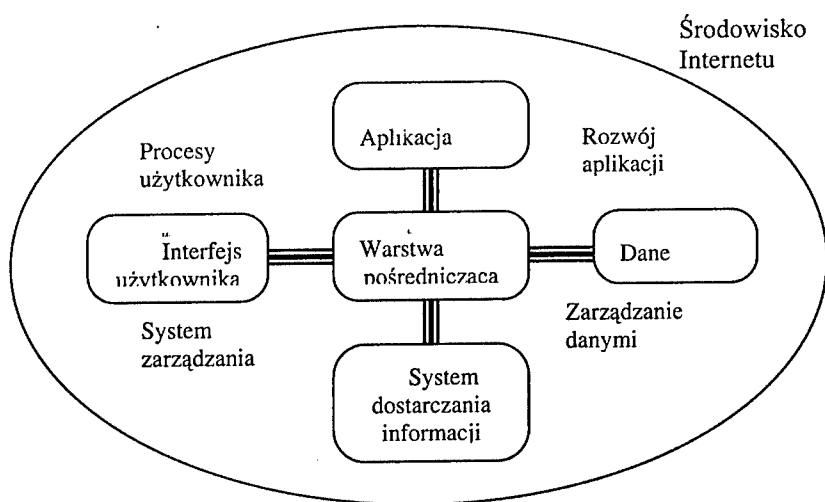
- ISP (*Information Service Provider*),
- ASP (*Application Service Provider*).

ISP – zapewnia dostęp do usług oferowanych przez sieć telekomunikacyjną, związanych z realizacją typowych procesów komunikacyjnych i ewentualnie wyszukiwania, czy zapamiętywania wymaganej informacji.

ASP – zapewnia dostęp do różnych rodzajów aplikacji od prostej witryny internetowej poprzez handel elektroniczny do planowania przedsiębiorstwa. Wykorzystuje się przy tym Internet, intranety lub ekstranety. Co więcej, ASP umożliwia (w oparciu o oferowane usługi) budowę nowych aplikacji bez ogromnego wysiłku i kosztów.

Koncepcja taka spaja na przykład tradycyjny outsourcing z tradycyjnym utrzymaniem sieci, umożliwia dostęp do odpowiedniego oprogramowania (payroll), zapewnia jego utrzymanie oraz wykonanie na sprzęcie dostawcy aplikacji. Poza tym ASP administruje siecią użytkownika i zapewnia utrzymanie (support) aplikacji wykorzystywanych przez daną organizację. Do oferowanych aplikacji przez ASP zalicza się:

- Internet i portale zewnętrzne WWW,
- hurtownie bazy danych,
- obsługa działów danej organizacji (komunikacja pionowa),
- handel, księgowość, listy płac,
- współpraca między organizacjami (komunikacja pozioma),
- planowanie zasobów przedsiębiorstwa,
- sprzedaż towarów i usług.



Rys. 5. Komponentowy model architektury Internetu

Wiele firm informatycznych oferuje określone aplikacje jako ASP. Są to tak znane firmy, jak: SAP, Oracle, IBM, czy niektóre firmy telekomunikacyjne. Z uwagi na opłacalność takich transakcji wiele różnych nieinformatycznych przedsiębiorstw jest zainteresowanych tego typu usługami i aplikacjami.

Własności konwergencji i synergii sprawiają, że przedstawiony poprzednio model warstwowy coraz bardziej nie odpowiada już rzeczywistości, gdyż w złożonych systemach, oprócz powiązań hierarchicznych może występować wiele istotnych relacji na tych samych poziomach abstrakcji. Przykładem tego jest właśnie system WWW, którego architekturę określa się jako zbiór różnych komponentów zanurzonych w środowisku Internetu (patrz rys. 5). Są to następujące komponenty:

- Interfejs użytkownika – odpowiedzialny za współpracę użytkowników z Internetem. Oprócz tradycyjnego interfejsu komputera GUI (*Graphical User Interface*) w jego skład wchodzi również „interpreter” języka definicji stron i przeglądarka HTML, pracująca pod Netscape czy MS Internet Explorer.
- Aplikacje – reprezentujące możliwości przetwarzania w zakresie programów, motorów czy agentów. Wykorzystują one interfejs API (*Application Program Interface*) dla zapewnienia współpracy z innymi komponentami tej aplikacji lub różnych systemów.
- Dane – stanowiące informacje i wiedzę zapamiętaną w rozmaitej formie, włączając w to strony HTML, odsyłacze, zbiory, wirtualne bazy danych, dokumenty, obiekty, dane multimedialne, hurtownie danych. Istotne są tutaj standardy języków zapytań, np. SQL, czy standardy przekazu i kompresji danych JPEG, MPEG, PDF.
- Systemy przetwarzania informacji – zawierające urządzenia przetwarzania, zapamiętywania i przesyłania informacji, takie jak: terminale użytkowników, serwery, routery. W skład tego komponentu wchodzi też: maszyny wirtualne np. języka JAVA (JVM – *Java Virtual Machine*) wraz z urządzeniami ją wykorzystującymi PVM (*Parallel Virtual Machine*), a także protokoły sieciowe, np. TCP/IP.

Warstwa pośrednicząca – będąca rozproszonym systemem operacyjnym włączającym protokoły połączeń dla czterech innych komponentów oraz protokoły komunikacji pomiędzy pracującymi procesami. Wykorzystuje się tutaj między innymi standard HTTP, standard usług CORBA, itp.

Internet funkcjonuje w pewnym środowisku, lub inaczej, do wykonania jego zadań niezbędne są cztery elementy tworzące środowisko zewnętrzne. Obejmuje ono:

- zarządzanie systemami – tj. monitorowanie, identyfikację sytuacji wyjątkowych, dostępność do mocy obliczeniowej, rutowanie i wspomaganie automatycznego zarządzania procesami webowymi,
- zarządzanie danymi – czyli współpracę z bazami danych i wiedzy, konfigurowanie danych, ustalanie standardu zapytań, wyszukiwanie semantyczne,
- rozwój aplikacji – wiążący się z budową oprogramowania, w tym: wytwarzania i dystrybucji apletów ponownie używanych, specyfikacją JVM, nowymi paradygmatami zachowania się końcowych użytkowników, uzupełnianiu COTS (*Commercial-Off The Shelf*) przez SOTI (*Software Off The Internet*), przejściem od „processor-driven tools” do „tool-driven processors”, różnymi modelami wytwarzania oprogramowania,
- procesy użytkownika – reprezentujące współpracę użytkownika z konkretną aplikacją, którą realizuje Internet.

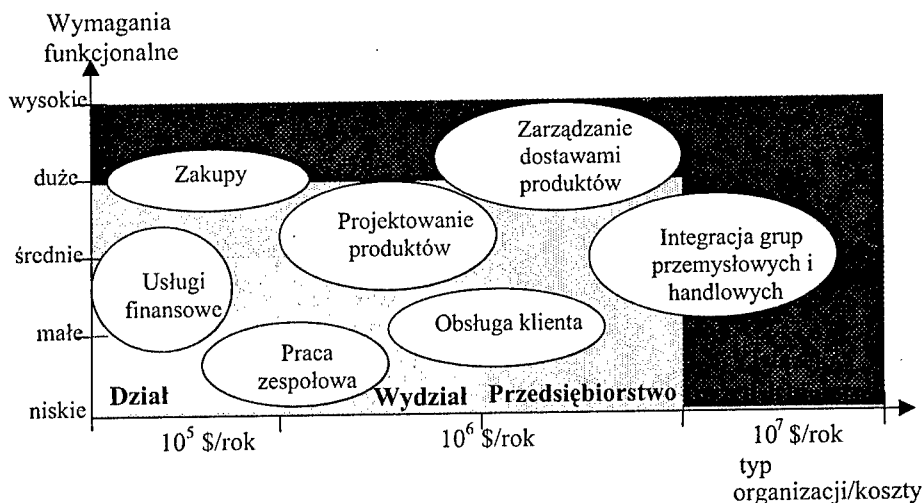
Te cztery elementy środowiska są ściśle związane z pięcioma komponentami Internetu. Warstwa pośrednicząca łączy je w jedną spójną całość – przez co jej znaczenie będzie stale wzrastać. Przykładem tego są takie rozwiązania jak: DCOM (Distributed Common Object Model) – firmy Microsoft, czy CORBA (*Common Request Broker Architecture*) – grupy firm OMG (*Object Management Group*). Wyłaniają się obecnie dwa podstawowe standardy .NET i J2EE.

Wytworzenie aplikacji wymaga opracowania modułów, jak też sprecyzowania zależności pomiędzy nimi, a pozostałymi komponentami systemu. Zatem metody projektowania sieciowych aplikacji informatycznych będą ciągle modyfikowane. Prześledzimy to na przykładzie rozwoju aplikacji ekstranetowych. Przyjmuje się, że już niedługo większość zadań związanych z przesyłaniem i przetwarzaniem informacji, zarówno wewnątrz przedsiębiorstw, jak i na zewnątrz będzie skomputeryzowana i usieciowiona. Możliwe są wtedy trzy rozwiązania:

1. Korzystanie z Internetu, który stosując odpowiednie standardy zapewnia globalny dostęp i wysoką niezawodność funkcjonowania.
2. Korzystanie z intranetu, który usprawnia wewnętrzną komunikację w całej firmie, dostarcza odpowiedniego bezpieczeństwa oraz umożliwia dalszą integrację systemów.
3. Korzystanie z ekstranetów, które umożliwiają wykreowanie globalnego rynku, na którym odpowiednie transakcje biznesowe zaangażują wiele różnych firm – często rozlokowanych w odległych geograficznie i ekonomicznie regionach.

Zakłada się, że firmy wykorzystujące technologię intranetów, udostępniają jednocześnie swoim bliskim partnerom biznesowym określone informacje, aplikacje i usługi. Dzięki temu intranety w ewolucyjny sposób przekształcają się w ekstranety, które z kolei stają się obecnie główną platformą handlu elektronicznego. To z kolei będzie stymulowało rozwój nowych aplikacji oraz narzędzi do ich rozwoju (np. wirtualnych magazynów, czy usług katalogowych). Na rys. 6 przedstawiono koncepcję rozwoju aplikacji ekstranetowych w najbliższych latach.

Przyszłe systemy informatyczne będą umożliwiały użytkownikowi dostęp do wymaganej przez niego informacji, bez względu na to, kiedy zgłosi takie żądanie, bądź gdzie aktualnie przebywa. Przykładem tego kierunku rozwoju są tzw. wirtualne systemy sieciowe VNC (*Virtual Computer Networks*). Komputery sieciowe (NC) to tanie urządzenia zapewniające prosty dostęp do zdecentralizowanych zasobów sieci. Pracują one jako klienci i odwołują się do znacznie mocniejszych (w sensie wydajności) komputerów-serwerów oferujących wykonanie programów, przechowywanie danych i wyników zgodnie z wymaganiami użytkownika. Wirtualne systemy sieciowe są rozszerzoną koncepcją polegającą na tym, że serwery wykonują nie tylko aplikacje, ale także kreują całe środowisko użytkownika, które jest dostępne poprzez Internet i wykorzystują symulatory prostych komputerów sieciowych. W przeciwieństwie do klasycznego Internetu, który zapewnia użytkownikowi dostęp do zasobów umieszczonych w dowolnym miejscu na świecie, z jego domowego środowiska obliczeniowego, wirtualne systemy sieciowe zapewniają dostęp do dowolnych zasobów z dowolnego punktu w świecie. Tak więc VNC umożliwi „przenośne” obliczenia bez wymogu zabierania ze sobą własnych urządzeń (komputerów biurowych), co więcej, pojawi się możliwość dostępu do własnego środowiska z wielu miejsc jednocześnie – co będzie bezpośrednio wspierać pewne przydatne formy współpracy.



Rys. 6. Rozwój aplikacji ekstranetowych w latach 1998 – 2002

Znamienny jest też dalszy rozwój systemów multimedialnych, który wraz z postępem technologicznym (elektroniki, optyki, genetyki) spowoduje nie tylko zwiększenie szybkości transmisji informacji (od 1 Gb/s w 1990 roku do co najmniej 100 Gb/s w 2010 roku) i szybkości jej przetwarzania, ale również zmieni same sposoby przetwarzania – w kierunku przetwarzania zespołowego (*collaborative computing*), w którym zdecydowaną rolę odgrywać będzie nie tylko nagromadzona wiedza, ale i mechanizmy jej wyszukiwania, a następnie odpowiednie wnioskowanie. Przykładem takiego podejścia może być współpraca przy budowie międzynarodowej stacji kosmicznej (ISS – *International Space Station*), której zadania rozłożone na 25 lat dotyczyć mają: badania kosmosu, zapewnienia miejsca

„tranzytowego” podróży na Marsa, przewidywania klimatu, itp. W tym celu planuje się rozwój inteligentnego środowiska ISE (*Intelligent Synthesis Environment*), którego zadaniem będzie „powiązanie” naukowców, zespołów projektowych, producentów, dostawców i konsultantów, nie tylko w celu opracowania tak złożonego systemu, ale również w celu ustalania jego różnych misji. To z kolei stworzy nie tylko nowe modele przetwarzania zespołowego, ale również modele zanurzania obliczeń (*immersive*), a także systemy zarządzania wiedzą.

5. ZAKOŃCZENIE

Konwergencja i synergia doprowadziły do rozwoju szerokiego wachlarza usług i aplikacji informacyjnych, bardzo istotnych dla funkcjonowania gospodarki i życia społecznego. Obecnie, z usług takich korzysta od około 15% – w krajach rozwiniętych do około 45% – w krajach wysoko rozwiniętych – społeczeństw tych krajów. Rodzi się jednak pytanie, w jaki sposób rozwój aplikacji informatycznych i poprawa jakości usług wpłynie na poziom życia społeczeństwa, i czy szeroka globalizacja i unifikacja systemów i usług będzie prowadziła do zatarcia różnic i cech specyficznych wielu społeczeństw. Pytanie to pozostaje jeszcze bez odpowiedzi, choć pojawia się zarówno wiele głosów pozytywnych, wychwalających postęp technologiczny, jak też sporo głosów krytycznych, sygnalizujących możliwe zagrożenia (problemy bezpieczeństwa, izolacji jednostek, itp.). Jest to sygnał, że konwergencja i synergia rozszerzają swoje oddziaływanie i kreują nowe pojęcia z pogranicza techniki i humanistyki, które będą decydowały o obliczu świata cyfrowego i jakości życia - tworzącego się społeczeństwa informacyjnego. Przyjmuje się przy tym, jako sprawę bezdyskusyjną, iż rola i miejsce człowieka w takim świecie powinny pozostać niezmiennione – tzn. człowiek powinien zajmować najważniejsze, centralne miejsce. To z myślą o człowieku i jego różnorodnych potrzebach należy rozwijać świat cyfrowy, korzystając z ogólnie dostępnych dobrodziejstw techniki. W konsekwencji konwergencja i synergia powinny podlegać „filtrowaniu”, by funkcjonowanie społeczeństwa informacyjnego nie było kojarzone z epoką marzeń o „szklanych domach”. W tym celu rozsądne jest też tworzenie interdyscyplinarnych zespołów badawczych, podejmujących się rozwiązywania istotnych problemów oraz efektywnie i skutecznie wykorzystujących możliwości konwergencji i synergii.

BIBLIOGRAFIA

- [1] Krawczyk H., Woźniak J.: IT – jak je wykorzystać? Potrzebna strategia działania. (Warunki rozwoju i efektywnego wykorzystania IT w województwie pomorskim. Referat na Konferencji pt. „ Czy sektor IT jest szansą rozwojową dla województwa pomorskiego”. Gdańsk, 17 kwietnia 2000). Pomorski Przegląd Gospodarczy, s. 52-57, Nr 3-4, 2000.
- [2] Kućmierz W.: Tendencje techniczne i ekonomiczne rozwoju technologii mikroelektronicznych i ich zastosowań. Opracowanie dla Departamentu Strategii Gospodarczej Ministra Gospodarki R.P. Warszawa, 1997.
- [3] Ruszczyk Z.: Internet w biznesie. Gdańsk, 1997.
- [4] Węglarz J. I inni: Cele i kierunki rozwoju społeczeństwa informacyjnego w Polsce. Opracowanie dla KBN i Ministra Łączności, Warszawa, 1999.
- [5] Węgrzyn S.: Nanosystemy informatyki. IiIS – PAN. Gliwice, 1999.

- [6] Czyżewski A.: Dźwięk cyfrowy. Akademicka Oficyna Wydawnicza EXIT. Warszawa, 1998.
- [7] Kostek B.: Soft Computing in Acoustics. Physica – Verlag, Heidelberg – Germany, 1999.
- [8] Kubale M.: Introduction to Computational Complexity and Algorithmic Graph Coloring. Wydawnictwo GTN, Gdańsk 1998.
- [9] Krawczyk H., Wiszniewski B.: Analysis and Testing of Distributed Software Applications. Baldock: RSP (J. Wiley), 1998.
- [10] Mrozowski M.: Guided Electromagnetic Waves – Properties and Analysis. RSP (J. Wiley), 1997.
- [11] Niedźwiecki M.: Identification of Time – Varying Processes. RSP (J. Wiley), 2000 r.
- [12] Woźniak J., Nowicki K.: Sieci LAN, MAN, WAN – Protokoły komunikacyjne. Fundacja Postępu Telekomunikacji, Kraków, wyd. II 2000.

SYNERGY AND CONVERGENCY STIMULATING FORCES FOR DEVELOPMENT OF MODERN INFORMATION TECHNOLOGY

Summary

Electronics, telecommunications and informatics are disciplines of science and technique that strongly interacting with each other introduce new values (synergy) and create a common universal vision of a digital world (convergency and globalisation). The paper presents a survey of current and future researches and implementation problems strictly connected with information technology. Achievements and development directions of modern electronics are given. Their impact on progress in telecommunications and computer science is also presented. Tendencies observed in telecommunications, including aspects of convergency of networking technologies and services are described. The paper also presents analysis of new effective computing models and architectures software products and information oriented services showing the essential synergy aspects arising in the digital word.

Péter Arató, Bálint Kiss, László Vajta, and Gábor Vámos

**Department of Control Engineering and Information Technology,
Budapest University of Technology and Economics**

FRAUDULENT CONSUMER BEHAVIOR ANALYSIS AND DETECTION FOR UTILITY COMPANIES

Abstract

Fraudulent or negligent behavior of consumers of Utility Companies (UC) results important financial losses. Therefore virtually all such companies have departments carrying out regular consumer control based on lists of hypothesized fraudulent customers (LHFC). These lists are created using expert knowledge accumulated by the company personnel during long years of experience and statistical analysis of previously detected frauds and consumer databases. Completeness and high hit rate is required in the case of LHFCs in order to maximize efficiency and to minimize cost. This paper describes a methodology followed in a project conducted jointly with a Hungarian UC aiming to produce LHFCs for "huge" consumers.

1. INTRODUCTION

Large public utility companies (UC) in the Central and Eastern European region spend considerable amount of money in order to detect fraudulent or negligent (i.e. not intentionally fraudulent) consumers. The same companies maintain large electronic databases, primarily for invoicing purposes, with various consumption specific data and some records of previously discovered fraudulent or negligent cases. In general, the detection of a fraudulent consumer makes necessary the control of a large number of decent clients, lowering the mutual trust, the prestige of the company, and the control efficiency.

It is a straightforward idea to try to use the information provided by the discovered cases in order to create a general profile of a fraudulent consumer and to try to match this profile with the clients' behavior encoded in the database. This would result higher hit rate among controlled consumers and would decrease the unit cost of a successful detection of frauds.

Our department conducts a joint project with a Hungarian UC. First, we had to study the possibility to use existing databases to identify potentially fraudulent consumers, and then such consumers had to be selected for control. The fraudulent behavior profile was constructed based on human expert knowledge provided by company employees, and using the data of previously discovered and proven fraudulent cases, if appropriate.

The profile itself consists mainly of hypotheses on a selected set of consumer attributes and their values. Then these values are fitted with the corresponding attributes of each client in the consumer database and clients with the closest value sets are selected for control.

The systematic control of the selected consumers based on simple profiling assumptions has shown a higher hit rate compared to the case of arbitrary selected clients even if one uses the existing attribute sets of the database. A detailed study may also show that some additional attributes not present in the database for the moment would result even higher hit rate. (Recall however that the set of additional attributes is limited by the legislation on protection of personal information). Our future cooperation with the company aims the expansion of the studies to other services and consumer populations, which allow us to test the service specificity of the already obtained profiles.

The remaining part of the paper is organized as follows. The next section gives the main steps of the methodology we follow during the realization of our project. Section 3 describes the database and Section 4 deals with the profile modeling issues. The current results of the project are described in Section 5 preceding some concluding remarks given in the Conclusions.

In this paper, we concentrate on consumers who are not individuals but private or public companies and institutions. Let us remark that any topic dealing with the legislation related to UCs and the ways of the pursuit of fraudulent consumers on court are beyond the scope of this paper.

2. METHODOLOGY OVERVIEW

Let us give the main phases of our study. Depending on the UC and in particular on its infrastructure, some steps may require different amount of work.

1. Data replication. This step consists of the replication of the databases available at the UC. Some companies may have several databases for several invoice systems (some of them being quite outdated) which are different for one geographic region to another or for one consumer group to another. The data is replicated in a single unified relational database, which can be queried using SQL (e.g. Oracle, MS Access, etc.). The replicated database has to be kept up to date, which needs its regular maintenance.
2. Data consistency check. The consistency of the unified database is checked. Actually, the data consistency requirements are translated into internal logical rules of the database, e.g. all consumers must have registered meter(s); all consumption invoice record must have an associated meter reading record; no meter reading record can be associated to a consumer waiting to be connected, etc. These checks can be realized by more or less complicated but standard SQL queries and also allow to verify the performance of the database maintenance personnel or that of the contractor if this activity is outsourced by the UC.
3. Model creation. The model aims to profile fraudulent consumers and serves as a basis for the generation of list of hypothesized fraudulent customers (LHFC). This is an iterative process, see also the related Section for details.
4. LHFC generation. The raw source for the list generation is the result of one or more queries in the replicated database such that the query checks the fit of all consumers against the profile of the virtual fraudulent consumer. Some post-processing of these results is needed in order to optimize control costs (e.g. sorting based on geographic

locations and metering type since the manipulation of some high tension meters requires special equipment and staff).

5. Physical control of the previously selected customers.
6. Hit rate evaluation. The hit rate is defined as the ratio (in percentages) of the detected frauds w.r.t. the total number of consumers on the LHFC.

Besides hit rate evaluation, UC companies are also interested in indicators like the infection rate (w.r.t. a geographical region or a specific group of consumers). The general infection rate (GIR) is the ratio (in percentages) between the number of fraudulent consumers and the total number of consumers of the UC. The GIR can be estimated by controlling an arbitrary selected set of consumers or can be determined exactly by exhaustive control where the hit rate equals the GIR.

A good evaluation of the hit rate (in percentages) of a LHFC is provided by its comparison to the GIR. If the hit rate of the LHFC is larger than the GIR, the profile used to generate the list is based on relevant hypotheses and quantitative model. If the hit rate is close to the GIR, the profile or the database is useless, since they contain no relevant information about the behavior of fraudulent consumers.

The following personnel were involved in the different phases of our study.

1. Human Expert: he or she holds possession of practical and/or theoretical knowledge about the problem domain. These experts are employees of the UC working for the fraud detection department and having long years of experiences. They are the main source of the human knowledge incorporated into the model.
2. Statistician: skilled in the classical statistical methods of data analysis. He or she participates in the model creation.
3. Knowledge Engineer: proficient in the creation of knowledge based probabilistic model (see Section 4.1).
4. Database engineer: he or she manages the database replication and runs queries.
5. Technicians: personnel carrying out the controls on the field. This activity may need special training for a given group of consumers (special tension level and meter types) or may be also outsourced.

3. DATABASE

The replicated UC consumer database is a relational database and hence consists of a set of tables. One table contains records of all consumers with their ID (primary key), name, legal settlement, consumer state (active, cancelled, waiting for connection, etc.), starting date of the utility service, tariff code, and for some cases, the company status (Inc, Ltd., etc.). A separate table contains records of meters and meter readings. More than one meter can be associated to one consumer. Invoices are also in a separate table and are created based on the meter readings and tariffs applied to the client. The relationships among tables are given in Figure 1.

It is worth to note that some tables change less frequently than others do. Typically, meter readings occur every month and invoices are sent to clients accordingly. This part of the database is considered to be *dynamic*. The table of consumers changes when new clients are coming and old clients are leaving. Immediate control of new clients is not desirable from a commercial point of view, therefore a less frequent update of this table can be envisaged and this table is considered as the *static* part of the database.

The database contains also records of previously detected fraudulent or negligent behavior on behalf of the consumer (not shown in Figure 1). The associated table contains records of consumers including the nature of fraud, cost, detection date, etc.

In the sequel, the fields (or columns) in the tables of the database are also referred to as variables, denoted by V_i . The domain of the variable V_i is given by V_i . These variables get values for a given record, i.e. for a given consumer, meter, invoice or fraud.

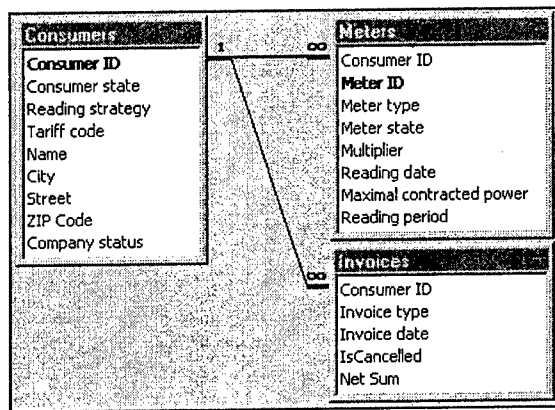


Fig.1. Relationships among tables in database

4. MODELING

The objective of the modeling is to find an abstract representation of the information coded in the database and that of the human knowledge accumulated as experience in the minds of UC employees.

Deterministic and probabilistic representations can be both considered. For deterministic representations, the model encodes a mapping or classification

$$M : fV_1;V_2;\dots;V_n \rightarrow \{F, \} \quad (4.1)$$

where F is a Boolean variable determining whether the given values of attributes represent a potentially fraudulent behavior or not, and M is the mapping defining the fraudulent behavior profile. Different types of "universal" approximators can be used to realize the mapping M , such as neural networks, fuzzy mappings, etc. Recall that neural networks have been already widely used to detect fraudulent financial transactions [1,2,3].

Another possibility is to use probabilistic models such that all variables are aleatory variables and each record in the database is a realization of these variables. Then the model is given by the joint probability distribution function (PDF) over all variables as

$$P : fV_1;V_2;\dots;V_n;F \rightarrow [0;1]. \quad (4.2)$$

Taking the marginal density function by fixing the values of V_1, V_2, \dots, V_n , one gets the probability of fraud for a given consumer, based on the model encoded by P . Note that the mapping P can be also realized by a "universal" approximator.

In both cases, the major issue is to fit the mappings P and M with data available in the database and in form of human knowledge. A drawback of the neural networks is that they are unsuitable to incorporate human knowledge intuitively, therefore we opted for another approximator, the Bayesian network.

4.1. Bayesian Networks

The background we use to construct knowledge based expert systems for the consumer modeling applications with uncertainty is provided by the theory of Bayesian networks [4]. These networks are used to model a given problem domain by means of a joint PDF over a set of probabilistic variables as given by (4.2).

Since these networks have a well-interpretable graphical (qualitative) component and a numerical (quantitative) component, they suit the cases where the existing knowledge in the domain comes partly from human experts and partly from databases. In fact, human expert knowledge can be coded using mainly the graphical component, whereas data extracted from databases are used to train the quantitative component.

Definition [Bayesian networks]: A Bayesian network over a set of variables $U = \{V_1, V_2, \dots, V_n\}$ ¹ consists of a graphical and a quantifying component:

1. *Graphical component*: directed acyclic graph: G . Each node in the graph represents a variable in U . The set of parents of a variable V (i.e. the nodes from which there is an arc pointing to V) is denoted by π_V .
2. *Quantifying component*: each variable V in U (i.e. each node in G) is quantified with a conditional probability distribution function denoted by $P(V|\pi_V)$.

The Bayesian network encodes a joint PDF over U : $P(U) = \prod_{i=1}^n P(V_i|\pi_{V_i})$.

Suppose now that for some variables of the network we are able to obtain either their deterministic value or their distribution and we wish to calculate the distribution of a given query variable V_q (e.g. F) knowing the evidences, i.e. $P(V_q|E)$, where $E \subset U$ is the set of evidence variables. This is called inference and corresponds to the

$$P(V_q|E) = \sum_{U \setminus \{V_q\}} \prod_{i=1}^n P(V_i|\pi_{V_i}) \quad (4.3)$$

marginalization where all $V \in E$ is replaced by the values corresponding to the evidence. The resulting probability distribution combines the expert knowledge encoded in the network and the collected evidence.

The use of (4.3) in the inference would mean the marginalisation for all values of all variables other than the query variable and the hard evidence variables in E . This would lead to high computational cost. Therefore the marginalization is carried out in a secondary structure called tree of clusters using the so called Probability Propagation in Tree of Clusters (PPTC) algorithm [5]. Roughly speaking, cluster trees are undirected, acyclic graphs whose nodes are clusters containing sets of the nodes from the original network. The rules of transformations leading from the original Bayesian network to a cluster tree ensure that both entities correspond to the same joint PDF.

¹ The set of variable may comprise F

4.2. Model creation steps

A major problem in our project was that a relatively small number of discovered fraudulent cases were available, hence the reliability of the data on which the fitting is based is limited. Recall that the controlled but honest consumer data would be also useful to train the Bayesian network.

Therefore we opted first to create test control lists (small LHFCs) for two reasons: *a.* to test hypotheses based on discussions with UC employees; *b.* to have more fraudulent cases for model fitting. Recall that the first objective aims to obtain qualitative information to create the graphical component and the second one helps to determine qualitatively the conditional PDFs associated to the nodes of the Bayesian network. This modeling procedure results an iteration of steps as illustrated in Figure 2.

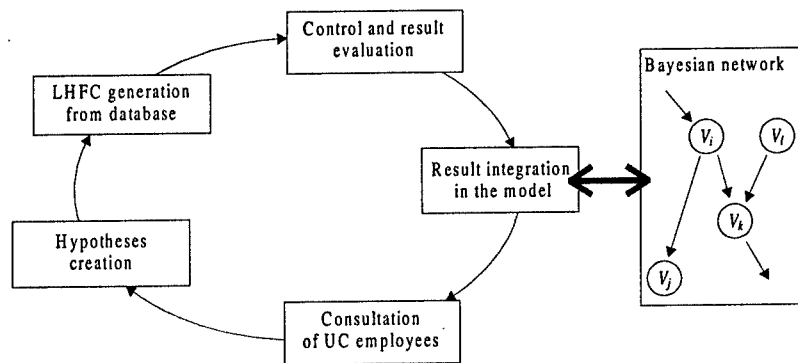


Fig.2. Iterative modeling

As an example, let us present an iteration aiming to study the dependence of the fraud rate on the company status and on the average net sum of invoices. Figure 3 shows the distribution of the yearly invoice net sum among consumers of the UC having Ltd. status.

A LHFC of 200 consumers is created selecting companies around the average yearly consumption value. All companies on the list are controlled by UC technicians. The obtained hit rate was 5.2%, which is considerably greater than the GIR estimated to be around 1%. It follows that the company status and the net sum of yearly invoices are to be included as variables in the Bayesian network. Note that approximately 2000 consumers were controlled in order to test different hypotheses.

5. CURRENT RESULTS

Currently, we finished the iterative modeling phase for the companies and for the institutional consumers. We got sufficient number of discovered fraudulent cases to carry out the training of the Bayesian network. Recall that the net sum, which is now estimated to flow onto the accounts of the UC due to the discovered frauds largely, overwhelms the cost of the whole project itself.

An initial phase is also started to obtain similar fraudulent behavior profile for individual consumers. The main difference w.r.t. company and institutional consumers is that the invoicing system is more complicated (trimestrial and yearly payments are

allowed) and the number of consumers is considerably higher. It follows that the LHFCs used to check working hypotheses in the iterative modeling process described in the previous section must contain more consumers in order to obtain statistically relevant data which means a higher cost for the model creation.

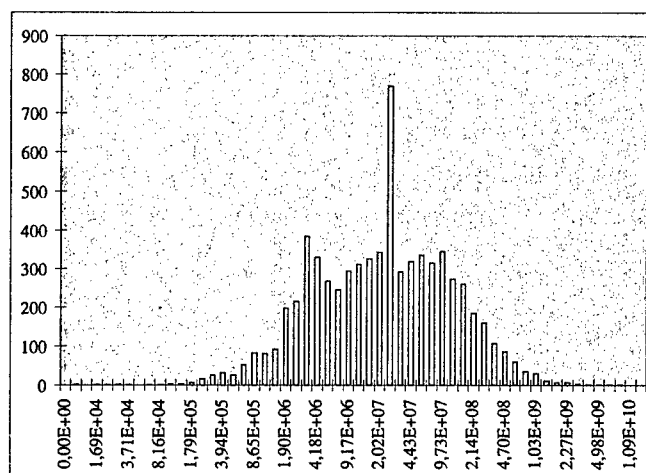


Fig.3. Invoice net sum distribution of Ltd.-s in the database

6. CONCLUSIONS

The paper discussed our project conducted to model fraudulent consumer behavior among clients of UCs. The profile can be represented by a Bayesian network which can be constructed and trained using iterative hypotheses testing based on human expert knowledge and statistical analysis of data available from replicated and unified UC databases.

A major issue remains to decide whether the variables present in the UC databases are sufficient to establish models generating LHFCs with high hit rate (i.e. at least higher than LHFCs produced using a single hypothesis). The study of this question needs future research work.

BIBLIOGRAPHY

- [1] Cerullo, M.J. and Cerullo, V.: *Using Neural Networks to Predict Financial Reporting Fraud*, Part 1-2. Computer Fraud & Security, p. 14-17., June-July 1999.
- [2] Richeson, L., Zimmermann, R.A., and Barnett, K.G.: *Predicting Customer Credit Performance: Can Neural Networks Outperform Traditional Statistical Methods?* International Journal of Applied Expert Systems, vol.2, no.2, p. 116-130., 1994
- [3] Dhar V. and Stein R.M.: *Neural Networks in Finance: The Importance of Methodology Over Technology*. PC AI, vol.12, no.3, p. 16-20., 1998.
- [4] Russel, S.J. and Norvig, P.: *Artificial Intelligence. A modern approach*, Prentice Hall, 1995.
- [5] Lauritzen, S.L. & Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their applications to expert systems. Journal of Royal Stat. Soc. **50**:157-224, 1988.

ANALIZA I WYKRYWANIE NIEWŁAŚCIWYCH ZACHOWAŃ KONSUMENCKICH W SEKTORZE PRZEDSIĘBIORSTW ŚWIADCZĄCYCH USŁUGI PUBLICZNE

Streszczenie

Celowo niewłaściwe lub niepożądane zachowania konsumenckie w sektorze przedsiębiorstw świadczących usługi publiczne powodują w efekcie znaczne wymierne straty finansowe. Dlatego też naturą rzeczy przedsiębiorstwa sektora publicznego posiadają wydzielone oddziały dla przeprowadzania regularnych kontroli zachowań konsumenckich w oparciu o szczegółowo opracowane listy klientów, którzy potencjalnie mogą zachowywać się celowo niewłaściwie. Listy takie tworzone są w oparciu o wiedzę ekspercką zbieraną przez pracowników tych przedsiębiorstw w wieloletnim okresie czasu i poparte są analizą statystyczną przypadków wykrytych już niewłaściwych zachowań w przeszłości oraz analizą baz danych konsumentów. Wysoki stopień wykrywania niewłaściwych zachowań jest bardzo pożądanym ze względu na możliwości uniknięcia zbędnych kosztów, a co za tym idzie zwiększenia efektywności przedsiębiorstwa. Artykuł przedstawia metodologię wykrywania konsumentów, którzy potencjalnie mogą wykazywać niewłaściwe zachowania. Metodologia rozwinięta została w formie praktycznego projektu aktualnie prowadzonego wraz z węgierskimi przedsiębiorstwami użyteczności publicznej w celu detekcji tych potencjalnie niewłaściwie zachowujących się konsumentów, którzy powodują największe straty.

Piotr Brudło

**Katedra Architektury Systemów Komputerowych,
Wydział Elektroniki Telekomunikacji i Informatyki, Politechnika Gdańska**

THE CONCEPT OF SMART AND SECURE LABORATORY

Summary

In the globalisation age, there is a necessity for development of structures providing solutions to problems which humanity faces. Academic institutions should adjust to the new situation and co-operate at the much higher level than before. Such co-operation, either bilateral or multilateral, should involve partners from both government and industry. This is because of the importance of finances, law, and technology as well the urgency to cope with global crises. Relationship between Poland and the United States is becoming of a growing importance because, due to new geopolitical circumstances, Poland is emerging as a new strategic partner for the United States in Europe. This situation provides new opportunities for co-operation among American and Polish academia.

1. INTRODUCTION

Globalisation and the American-Polish relationship call for a new framework for the co-operation between the Gdańsk University of Technology, Poland [1] and the University of New Hampshire, USA [2]. It should be noted that such informal co-operation between the two institutions has been existing for 20 years and included mutual visits, seminars, student exchanges, etc.

1.1. Mission

It is proposed to establish a virtual organisation between the Gdańsk University of Technology and the University of New Hampshire with focuses on:

- identification and solutions for critical problems concerning homeland security, particularly terrorism and health threats;
- identification and solutions for critical problems concerning public safety;
- identification of network of partners from government and industry to address challenges in homeland security and public safety and to establish a multifaceted team for problem solving;

- rapid prototyping and dissemination of technological solutions to be verified in lab and field environments and, in a case of success, to be deployed as quickly as possible;
- new American-Polish companies established through incubation and commercial parks initiatives.

In general, the above objectives are technology transfer and knowledge technology transfer focused on homeland security and public safety issues.

1.2. Infrastructure

The organisational structure of the international virtual organisation is presented in figure 1.

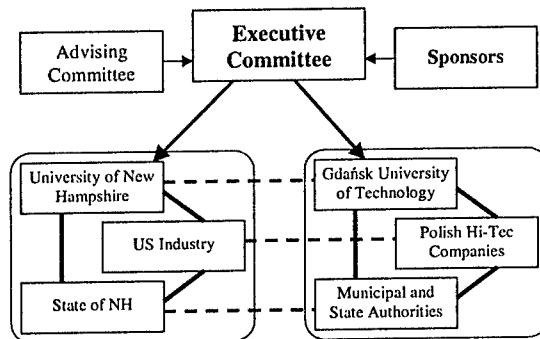


Fig. 1. Organisational structure of the international virtual organisation

It should be noted that even though the nucleus of the organisation is represented by the two academic institutions, other academic organisations, regardless of their origin and providing that they comply with the spirit and bylaws of this organisation, are welcome to join. Proposed organisation is organic in nature, i.e. new projects and new ideas will be evaluated and considered. The current focus is in the area of the homeland security and public safety technologies.

2. PROJECT PROPOSITION

A key component of the proposed project is to establish a smart and secure research laboratory (S²Lab). The smartness would entail heterogeneous networks of experimental and non-experimental sensors for a multiplicity of detection: fire, structure fatigue, earthquake, hazardous material spills, anthrax, small pox, etc. The security will entail the creation of denial area around the lab, monitoring systems, and the establishment of autonomous computer networks for emulating a hostile environments for hacking and virus defences. The lab would serve as educational ground for students working on security engineering project. As part of education, the students may study and design partial subsystems for a smart and secure research laboratory. The inadequacy of current research facilities from a security standpoint is generally well known. A laboratory of this new type requires multidisciplinary effort, for example electrical engineering students may design and implement VLSI crypto devices, biometrics, and sensors; computer science students

may address the issue of viruses and network security; chemical engineering students investigate technologies discouraging unauthorised persons from entering the research lab area; mechanical engineering students could design mechanical systems of such laboratory; civil engineering students could investigate the robustness of the laboratory against natural e.g. earthquake and man-made disasters e.g. terrorist attack.

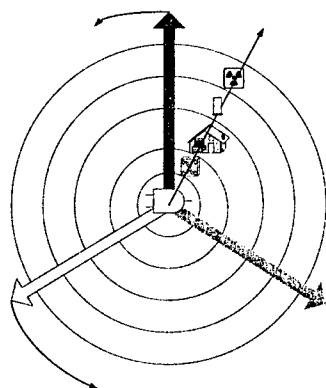


Fig. 2. Stationary part of the smart and secure laboratory

The proposed lab has several distinctive zones (figure 2):

The Denial Zone – the area surrounding the building or its section with a limited access. Students from all departments are encouraged in participating in projects pertinent to this layer, but especially students from Chemical Engineering, Chemistry, and Physics. We are interested in innovative concepts for denial of access to or occupation of an area by vehicles or personnel.

The Security Zone – this is a protective entry/exit zone, a gateway to the lab. Students from all disciplines are invited to participate in activities at this level. Project may involve biometrics, monitoring, smart cards etc.

The Lab Layout – this level represent a physical structure of the lab. Students from mechanical, chemical, and civil engineering are especially encouraged to conduct design studies of the lab layout.


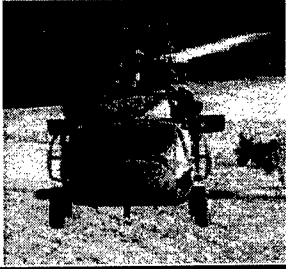
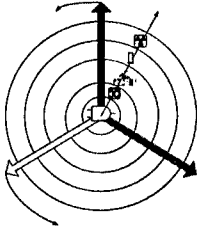
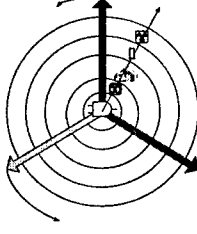
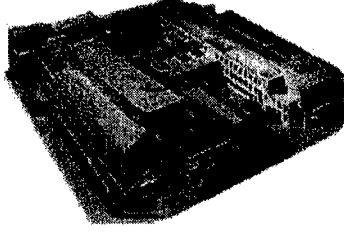
The Smart & Secure Network – it is envisioned that walls, floors, and ceilings of the lab will be “smart” e.g. saturated with systems of sensors for detection of chemicals, vibration, shocks, temperature, weather conditions, fire, and others. Those sensors will be connected through either fibre optics or wireless network to allow monitoring of the lab and detect undesired conditions. This part of the project will be conducted in co-operation with industry. Especially computer science and electrical and computer engineering majors are encouraged to participate in designing and programming this class of networks. Another area with this category is a computer network that can be used as a test bed for developing protection against hackers, trojans, and viruses.

The Crypto Microelectronics – specific VLSI or FPGA designs are considered at this level to demonstrate security concepts using the DES algorithm. Only students from the Department of Electrical and Computer Engineering have sufficient background to be involved in these designs.

The S²Lab itself will become a test-bed for the physical protection of a specific area. Students will be able to work on technologies for personal identification and verification (facial, voice, eye, fingerprints, etc.), combination and coding, individual tracking, etc.

Table 1

Scheme of the dual smart and secure research laboratory

	Gdańsk University of Technology	University of New Hampshire
M o b i l e p a r t		
S t a t i o n a r y		
L o c a t i o n	Two levels in a brand new building <i>In construction</i>	

The S²Lab projects can be local or remote and involve:

- monitoring structural fatigue (buildings, bridges, roads),
- monitoring the environment inside and outside buildings,
- determining energy efficiency,
- detecting pollutants and harmful chemicals,
- authenticating and tracking individuals or vehicles,
- anti-tampering techniques,
- computer/network security,
- etc.

Topics:

- homeland security – detection – prevention – rapid response,
- environmental monitoring, – air/water quality – energy efficiency,
- security – physical – web/DB – information systems,

- smart highways – road conditions – classification of vehicles,
- structures – smart walls – bridge monitoring – building monitoring,
- anti tampering – cargo containers – airport baggage – food chain area denial – people – vehicles.

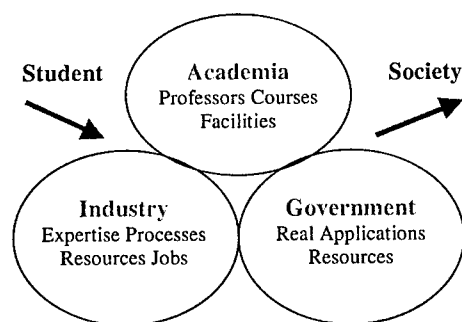


Fig. 3. Circles of interest for the smart and secure laboratory

The smart and secure laboratory will provide students with:

- multidisciplinary education,
- a chance to solve real world problems e.g. homeland security,
- a chance to work with industry and government.

The university with:

- state-of-the-art test-bed and learning facility,
- a showcase for a new "smart building" paradigm.

3. S²LAB – A CENTRE OF EXCELLENCE

Centres of excellence are units or organisational structures involved in scientific research and the development of high technology at a world level in terms of measurable scientific effects (including training activity).

These centres serve as hubs for teams of scientists with outstanding achievements who co-operate in the areas of common interest and of large significance for national economies. This role supports the innovative activity of the centre and boosts the promotion of research, technology and products at home and abroad.

Researchers working in the centres of excellence focus on strategic problems, using the scientific infrastructure of several institutions collaborating under one scientific and organisational management, while preserving a relatively high degree of autonomy.

In their programmes, centres of excellence do not set out to create new research institutions but rather to become a certain kind of "laboratories" co-operating actively with the industry and other research users. Centres of excellence should implement projects in fundamental research, as well as search for specific innovative applications. The size of the team and the available research infrastructure have to be sufficient to complete the scheduled work.

4. CONCLUSIONS

Security is probably the most important issue facing the humanity today. Despite geopolitical distance between Poland and the United States, there exists a remarkable amount of similarities in this field. In both countries, scientists and researches may and should contribute to the increase of security level by developing new technologies and relevant applications. It is proposed to develop an open-ended multidisciplinary co-operation scheme between the Gdańsk University of Technology and the University of New Hampshire in security engineering. It is proposed to establish a virtual security engineering laboratory at both sites that may serve as a test bed for new technologies before they are implemented in the field.

BIBLIOGRAPHY

- [1] <http://www.pg.gda.pl/>
- [2] <http://www.unh.edu/>
- [3] <http://www.kbn.gov.pl/>
- [4] <http://www.bbn.gov.pl/>
- [5] <http://www.whitehouse.gov/homeland/>

WIRTUALNE LABORATORIUM BEZPIECZEŃSTWA

Streszczenie

W chwili obecnej, w erze globalizacji istnieje niezbędna konieczność rozwoju struktur gwarantujących zapewnienie bezpieczeństwa dla zasadniczych wyzwań stawianych przed ludzkością. Stąd potrzeba ujęcia aspektu bezpieczeństwa w kategoriach globalnych. Instytucje akademickie dostosowują się w tym zakresie stosunkowo szybko i zaczynają wyznaczać nowe trendy oraz prowadzą współpracę z instytucjami pozaakademickimi. Tego typu kooperacja, zwykle wielostronna, włącza w swoje szeregi zarówno partnerów rządowych, samorządowych, instytucjonalnych jak i przemysłowych. Spowodowane jest to istotnością całościowego ujęcia aspektów finansowych, prawnych, technologicznych oraz wspólnym przewidywaniem sytuacji wyjątkowych i kryzysów. Współpraca pomiędzy Polska i USA w tym zakresie w chwili aktualnej zaczyna odgrywać coraz większe znaczenie, zarówno z powodów geopolitycznych, jak i z przyczyny ogromnego zaangażowania Stanów Zjednoczonych na obszarze strategicznego partnera w Europie Wschodniej. Sytuacja taka stworzyła ogromne i niepowtarzalne możliwości kooperacji pomiędzy akademickimi instytucjami ze strony polskiej i amerykańskiej. W tym aspekcie pojawiła się idea głębokiej i wielopłaszczyznowej współpracy pomiędzy Politechniką Gdańską oraz Uniwersytetem w New Hampshire, USA. Należy w tym miejscu nadmienić stałość współpracy prowadzonej pomiędzy powyższymi uniwersytetami od ponad 20 lat oraz duże wzajemne osiągnięcia wyrażające się zarówno ilością wspólnych projektów i opracowań jak i nowatorskich rozwiązań.

Grzegorz Górski

Katedry Systemów Informacyjnych, Politechnika Gdańska

BEZPIECZNE SIECI VLAN Z AUTORYZACJĄ DOSTĘPU OPARTE NA CERTYFIKATACH ATRYBUTÓW

Streszczenie

W referacie przedstawiono protokół dostępu do zasobów zrealizowany przy wykorzystaniu dynamicznej sieci VLAN z autoryzacją dostępu. Proces autoryzacyjny jest procedurą, która pozwala na weryfikację tożsamości i walidację żądania obsługi. Artykuł zawiera prezentację działania algorytmu sieci VLAN, w którym autoryzacja dostępu do zasobu odbywa się w oparciu o tzw. certyfikaty atrybutów, szczegółowo zdefiniowane w zestawie standardów X.509 – zalecanych przez ITU. Jest to nowatorskie rozwiązanie, które może stać się podstawą do budowy niezależnych od platformy sprzętowej i programowej, rozproszonych systemów autoryzacji dostępu do zasobów i usług w sieciach korporacyjnych.

1. WSTĘP

Wśród implementowanych i wykorzystywanych w praktyce sieci VLAN najczęściej spotyka się tzw. wirtualne grupy statyczne. Są one tworzone w trybie off-line przez administratora, który przydziela wybraną stację roboczą do danej grupy wirtualnej. Opierając się najczęściej na informacji o adresie stacji, która zawarta jest w napływających od stacji nagłówkach pakietów, tworzona jest lista grup roboczych, do których powinna należeć dana stacja robocza. Dynamiczne grupy wirtualne umożliwiają zmianę liczby członków sieci VLAN w trakcie pracy [1]. Powszechny dostęp do Internetu sprawił, że gwałtownie wzrosło zapotrzebowanie na bezpieczne metody wymiany informacji oraz dostępu do zasobów w sieciach korporacyjnych. W dynamicznej sieci VLAN zakłada się, że zbiór stacji roboczych w grupie może się zmieniać w czasie. Z tego powodu algorytmy autoryzacji dynamicznych sieci VLAN muszą dodatkowo obsługiwać procesy dołączania i odłączania stacji od sieci. Protokoły te w stanie „ustalonym” nie wymagają żadnej interwencji administratora. Przynależność do wybranej, dynamicznej grupy może zależeć od typu aplikacji, usług uruchamianych przez użytkownika (np. usługa poczty internetowej dostępna pod określonym standardowo portem protokołu TCP) lub identyfikatora użytkownika. Algorytmy bezpiecznych, dynamicznych sieci VLAN, które są tematem niniejszej pracy, stworzono na podstawie sieci z autoryzacją dostępu [2]. Proces autoryzacji najczęściej oparty jest się na jednym z trzech kryteriów tzn.: porcie fizycznym, regule logicznej lub identyfikatorze użytkownika. Weryfikacja użytkownika oparta o fizyczny

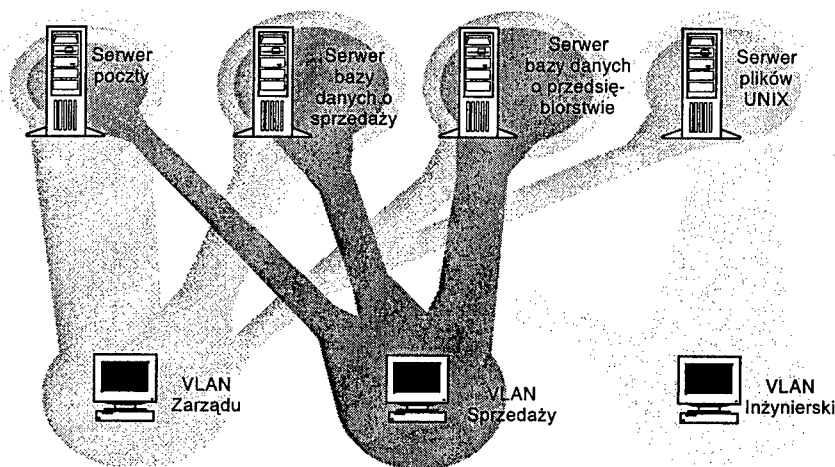
port zakłada, że administrator definiuje wybrany port przełącznika jako zaufany dla wybranej wirtualnej grupy roboczej. Na tej podstawie cały ruch pakietów pochodzący z tego portu jest uznawany za autoryzowany. Oczywiście to proste rozwiązanie nie zapewnia wymaganego poziomu bezpieczeństwa. Niedogodność tą usunięto w algorytmie z regułą logiczną, w którym port przełącznika jest powiązany z listą adresów sieciowych użytkowników. Tylko koniunkcja portu przełącznika z odpowiednim adresem sieciowym użytkownika zapewnia poprawną autoryzację stacji roboczej. Identyfikator użytkownika to kolejne kryterium, na podstawie którego może być przeprowadzony proces autoryzacji, bezpieczniejszy w porównaniu do dwóch wymienionych powyżej, autoryzacji. Stworzenie bezpiecznego algorytmu dynamicznej sieci VLAN jest możliwe dopiero po zastosowaniu infrastruktury klucza publicznego. Przedstawienie takiego algorytmu należy jednak poprzedzić opisem znanych sposobów organizacji dostępu do zasobów we współczesnych sieciach korporacyjnych.

2. ORGANIZACJA DOSTĘPU DO ZASOBÓW W SIECIACH KORPORACYJNYCH

Analiza dostępnych obecnie technologii, protokołów oraz aplikacji pozwala sformułować wniosek, że docelowym rozwiązaniem w administracji zasobami sieci korporacyjnych pracujących w środowisku heterogenicznym będą wirtualne grupy robocze VLAN [3], pozwalające na konsolidację użytkowników mających dostęp do wybranych usług czy zasobów. Zadania administratora rozproszonej sieci wykorzystującej technologię tego typu polegałyby na wyborze właściwego typu sieci wirtualnej tzn. ustaleniu kryterium przynależności, stworzeniu odpowiedniej liczby grup roboczych (VLAN) łączących użytkowników korzystających z wybranego zasobu a w końcu przypisanie kont użytkowników do sieci VLAN. W fazie projektowania struktury wirtualnych grup roboczych najczęściej wykorzystuje się dwa podejścia tzn.: model organizacyjny lub model uwzględniający oferowane w sieci usługi. Przy tworzeniu sieci VLAN według ustalonego kryterium przynależności, można wykorzystywać, podobnie jak w technikach przełączania pakietów, informacje z warstw łącza danych, sieciowej, sesji czy nawet aplikacji.

2.1. Model organizacyjny

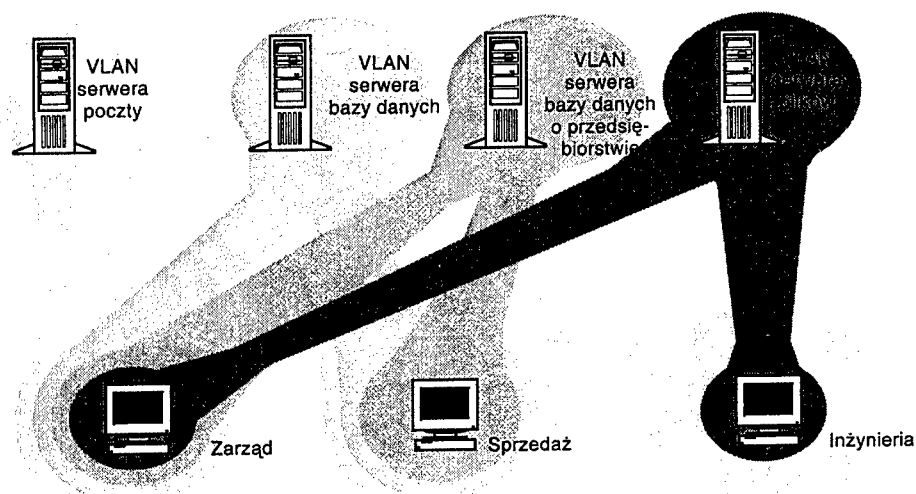
Organizacja przedsiębiorstw opiera się zazwyczaj na istnieniu pewnych grup funkcjonalnych (sekcje, działy, grupy robocze, itp.). W modelu tym każda z grup tworzy swoją sieć VLAN. Centralne zasoby, takie jak np. serwer poczty, należą do każdej sieci wirtualnej, a zatem wszyscy pracownicy mają dostęp do wybranych zasobów. Prawa dostępu do usług i aplikacji wykorzystywanych tylko przez wybraną grupę pracowników są nadawane poprzez właściwą sieć VLAN. Najważniejszą zaletą tego modelu jest jego prostota, a metoda zarządzania bardzo przypomina wykonywane od wielu lat czynności administrowania systemami operacyjnymi tzn. np.: nadawanie praw dostępu do plików i katalogów poprzez grupy robocze. Wykorzystanie tego podejścia wymaga jednak zwykle, aby tworzone grupy wirtualne były obiektami statycznymi tzn. tworzonymi i modyfikowanymi tylko przez administratora. Projekt sieci wykorzystującej takie rozwiązanie przedstawiono na rysunku 1.



Rys.1. Tworzenie sieci wirtualnych w oparciu o model organizacyjny

2.2 Model organizacyjny uwzględniający usługi oferowane w sieci

Podejście uwzględniające dostępne w sieci usługi, wymaga przedstawienia sieci komputerowej widzianej od strony pojedynczego użytkownika. Dla każdej aplikacji lub usługi sieciowej tworzona jest oddzielna sieć VLAN, do której, w zależności od potrzeb i uprawnień, dynamicznie podłączają się użytkownicy. Dla przykładu proces wysyłania listu przy pomocy poczty elektronicznej implikowałby chwilową przynależność do sieci VLAN zorganizowanej „wokół” serwera poczty. Rozwiązanie wykorzystujące model usług oferowanych w sieci przedstawiono na rysunku 2. Implementacja opisywanego modelu prowadzi jednak do stworzenia dużego zbioru sieci wirtualnych o bardzo skomplikowanej strukturze. Zarządzanie takimi sieciami wirtualnymi zmniejsza, co prawda nakład pracy administratora, lecz wymaga implementacji w przełącznikach skomplikowanych algorytmów analizujących napływające od użytkowników żądania obsługi. Wydajność tego typu sieci – mierzona dla przykładu opóźnieniem w dostępie do zasobu – będzie na pewno niższa niż „statycznych” sieci VLAN. Główną zaletą implementacji algorytmów opartych na modelu usług jest możliwość zastosowania dynamicznych sieci VLAN. W tego typu podsieciach wirtualnych decyzja o przynależności do wybranej domeny rozgłoszeniowej może zostać pozostawiona indywidualnemu użytkownikowi sieci. Sieć komputerowa ma charakter „dynamiczny”, definiowany w oparciu np. o adres multicastowy IP, na podstawie którego każda stacja ma możliwość wyboru dowolnego kanału obsługi.



Rys.2. Tworzenie sieci wirtualnych związanych z udostępnionymi usługami i zasobami

3. IMPLEMENTACJA CERTYFIKATÓW ATRYBUTÓW W PROTOKOLE DYNAMICZNEJ SIECI VLAN

Zastosowanie infrastruktury klucza publicznego PKI do uwierzytelniania członków grupy wirtualnej pozwala na stworzenie algorytmu, zgodnie z którym podłączający się do zasobu sieciowego użytkownik nie musi posiadać żadnej wiedzy np. identyfikatora i hasła przed rozpoczęciem procedury autoryzacji. Jedynym wymaganiem jest wyposażenie wszystkich użytkowników oraz rozproszonych komponentów protokołu VLAN, tzn. agentów i serwerów autoryzacyjnych, w aktywne certyfikaty zgodne ze standardem X.509 [4]. W prezentowanym protokole certyfikaty te są wykorzystywane do weryfikacji tożsamości osoby zgłaszającej żądanie obsługi. Proces taki rozpoczyna komponent protokołu określany mianem klienta autoryzacyjnego zainstalowany w stacji roboczej klienta. Weryfikacja tożsamości należy do najważniejszych zadań agenta autoryzacyjnego zaimplementowanego np. w przełączniku sieciowym, który komunikuje się z klientem. Napływające od klienta żądanie obsługi zawiera certyfikat klucza publicznego CKP nowego użytkownika. Jednym z pierwszych zadań agenta jest sprawdzenie statusu otrzymanego certyfikatu CKP w oparciu o publikowane przez Centrum Autoryzacji, które wystawiło certyfikat, listy certyfikatów unieważnionych CRL lub przy użyciu protokołu OCSP (ang. *On-line Certificate Status Protocol*). Po wykonaniu weryfikacji statusu certyfikatu agent generuje klucz sesji oraz hasło, które zostaje zaszyfrowane kluczem publicznym posiadacza certyfikatu (klucz ten jest częścią przesyłanej struktury certyfikatu). Tylko osoba posiadająca komplementarny klucz prywatny będzie mogła zdekodować wiadomość i przesłać do agenta ustalone hasło zaszyfrowane kluczem sesji. Fakt otrzymania poprawnego hasła i klucza sesji jest równoważna z weryfikacją tożsamości użytkownika. Wykorzystanie mechanizmów PKI zostało ze względów wydajnościowych ograniczone do bezpiecznego przekazania

symetrycznych kluczy sesyjnych. Dalszy wzrost poziomu bezpieczeństwa przesyłanych w grupie danych można osiągnąć okresowo wymieniając klucze sesji. Tak skonstruowany algorytm umożliwia tworzenie dynamicznych wirtualnych grup roboczych, które dzięki wykorzystaniu szyfrowania z kluczami symetrycznymi mogą poufnie wymieniać informacje w sieci korporacyjnej. Dla poprawnego działania protokołu należy jeszcze zdefiniować listę użytkowników, którzy mogą być członkami wybranej grupy wirtualnej. Taki zbiór informacji musi posiadać każdy agent. Przy tworzeniu dużych sieci VLAN jest to kłopotliwe pod względem administracyjnym oraz osłabia bezpieczeństwo protokołu. Udany atak na dowolnego agenta w sieci umożliwia nieautoryzowany dostęp do wszystkich zasobów sieci. Ponadto powstaje także problem czasowego dostępu do zasobu lub członkostwa w grupie wirtualnej np. na czas wybranej wideokonferencji.

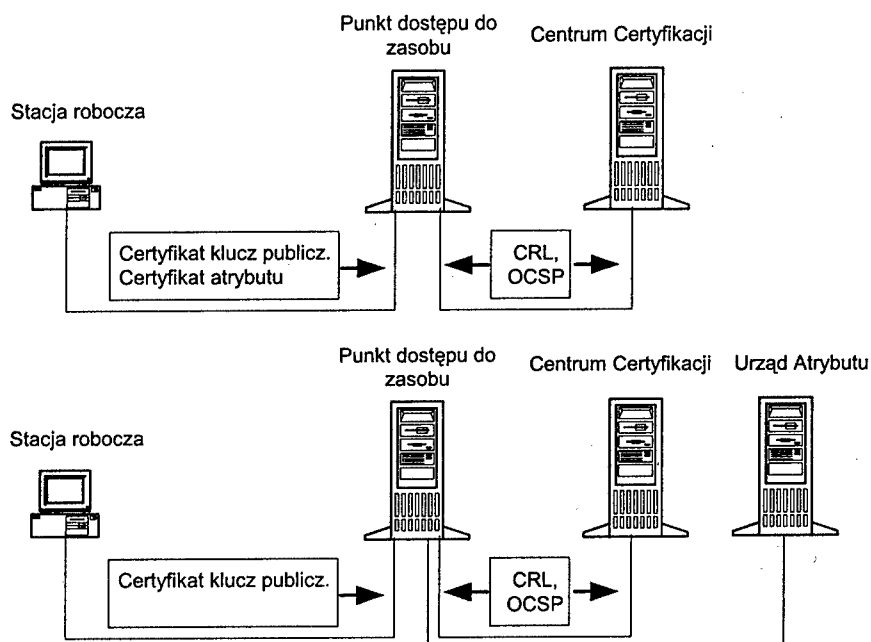
Wykorzystanie dodatkowych pól struktury certyfikatu klucza publicznego do określania listy zasobów i usług sieci dostępnych dla jego posiadacza nie jest zalecanym rozwiązaniem. Każda zmiana liczby lub nazwy zasobu skutkowałaby unieważnieniem certyfikatu, a zgodnie ze stosowanymi przez Centra Certyfikacji politykami bezpieczeństwa, wystawienie nowego certyfikatu wymaga osobistej weryfikacji posiadacza na podstawie innych dokumentów tożsamości. Bardzo wiele korporacji z przyczyn ekonomicznych nie posiada własnych centrów emitujących certyfikaty klucza publicznego, a korzysta z zewnętrznych w formie outsourcing'u, umożliwiając jednocześnie korzystanie z certyfikatów poza intranetem.

Opisane powyżej zadania są czasami określane jako system zarządzania przywilejami w sieci PMS (ang. *Privilege Managenemnt System*) [5], które możliwe są dzięki rozszerzeniom standardu X.509 o tzw. certyfikaty atrybutu. Obecne trwają dopiero wstępne prace standaryzacyjne nad certyfikatami warunkującymi dostęp do zasobu sieci (ang. *use condition certificate*). Do czasu ich opublikowania systemy PMS oparte są na certyfikatach atrybutu. W przeciwieństwie do Centrów Certyfikacji Klucza publicznego, certyfikaty atrybutu są wydawane przez Urzędy Atrybutu, które są najczęściej wewnętrzną usługą w sieci korporacyjnej. Certyfikat atrybutu jest strukturą zawierającą na początku numer i wystawcę certyfikatu klucza publicznego, które są referencją do struktury certyfikatu klucza publicznego, a następnie pozostałe pola określające nazwy oraz lokalizacje usług i zasobów, do których posiadać certyfikatu atrybutu ma dostęp. Przy poszczególnych nazwach zasobów może być dodatkowo określony przedział czasu, w którym dostęp jest możliwy, tak by po upływie czasu ważności dostępu dla jednego zasobu nie trzeba było unieważniać całego certyfikatu atrybutu. Standard X.509 [6] nie określa rozmiaru struktury certyfikatu atrybutu, a więc i liczby zasobów, które można do niego wpisać. Istotny jest natomiast mechanizm weryfikacji ważności takiego certyfikatu, którym jest podpis elektroniczny złożony pod strukturą, przy użyciu klucza prywatnego, przez administratora sieci lub inspektora bezpieczeństwa odpowiedzialnego za politykę dostępu do zasobów.

Pomimo, że standard nie zabrania organizowania dostępu do pojedynczych plików przy użyciu certyfikatów atrybutu nie jest to zalecane. Dużo efektywniej jest wykorzystać system PMS do kontroli dostępu do usług takich jak np.: konta pocztowe, katalogi, członkostwo w grupie wirtualnej. Certyfikaty atrybutu umożliwiają ponadto ustanowienie czasowego zastępstwa pomiędzy pracownikami, co jest szczególnie ważne w strukturze dużej korporacji. Wymaga to modyfikacji certyfikatu atrybutu pracownika zastępującego poprzez dopisanie nowych zasobów wraz z przedziałem czasu, w którym dostęp ma być przyznany. Nie zalecane, chociaż możliwe, jest wskazanie w certyfikacie atrybutu dodatkowej referencji (numer i wystawca) na certyfikat klucza publicznego pracownika

nieobecnego, gdyż w ten sposób daje się dostęp do absolutnie wszystkich praw włącznie z np. kontem pocztowym.

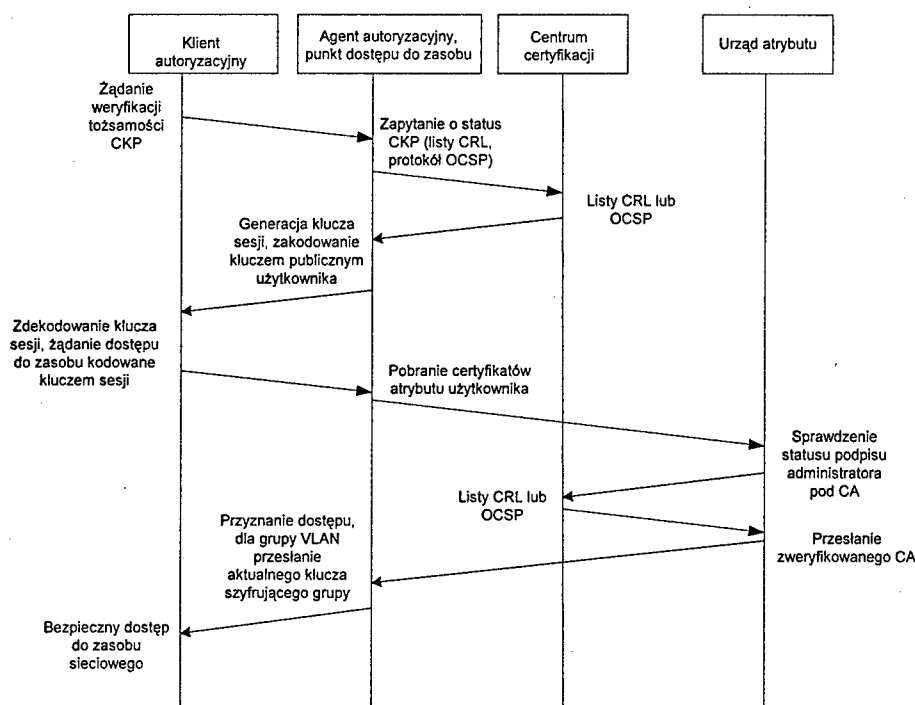
Praktyczną realizację systemu PMS, w oparciu o certyfikaty atrybutu, przedstawia rysunek 3.



Rys.3. Dwie wersje realizacji dostępu do zasobu z wykorzystaniem certyfikatu atrybutu

Autoryzacja dostępu do zasobu w oparciu o certyfikaty atrybutu może być przeprowadzona dwoma sposobami. Pierwszy z nich zakłada, że klient wraz z żądaniem obsługi dostępu do usługi, zasobu podłączenia do grupy VLAN przesyła swój certyfikat klucza publicznego oraz certyfikat atrybutu. Początkowym etapem autoryzacji jest opisana już weryfikacja tożsamości na podstawie certyfikatu klucza publicznego. Pomyślne zakończenie tego etapu rozpoczyna dopiero sprawdzanie przywileju dostępu. W tym celu sprawdzany jest certyfikat atrybutu nadesłany przez klienta. Ważność podpisu certyfikatu sprawdzana jest przy użyciu klucza publicznego inspektora bezpieczeństwa i oznacza ważność całej struktury. Punkt dostępu do zasobu np. serwer plików lub agent autoryzacyjny dynamicznej grupy VLAN udziela użytkownikowi dostępu do zasobu zawartego w żądaniu obsługi, tylko w przypadku, jeżeli zasób ten jest wymieniony w jego certyfikacie atrybutu. W zależności rodzaju informacji, która ma być przekazana klientowi może być ona przesyłana bezpiecznym kanałem z wykorzystaniem symetrycznego klucza szyfrującego sesji ustalonego w procesie weryfikacji tożsamości lub w sposób jawny bez szyfrowania. Drugi sposób autoryzacji jest identyczny pod względem rodzaju i kolejności przetwarzanych informacji, w tym przypadku żądający dostępu klient wysyła tylko certyfikat klucza publicznego oraz nazwę zasobu, natomiast certyfikat atrybutu jest

pobierany przez serwer z dostępnego w sieci korporacyjnej Urzędu Atrybutu. Pełny proces autoryzacji dostępu użytkownika do zasobu sieci pokazano w postaci diagramu na rysunku 4.



Rys.4. Proces autoryzacji dostępu do zasobu – wersja druga

W przypadku podłączania się użytkownika do dynamicznej grupy wirtualnej zamiast serwera posiadającego wybrany zasób sieciowy klient komunikuje się z agentem autoryzacyjnym, którego rola została już wcześniej opisana na etapie weryfikacji tożsamości. Agent autoryzacyjny przekazuje zamiast losowo wygenerowanego klucza sesji, klucz szyfrujący używany w danym momencie do kodowania informacji w wybranej, dynamicznej grupie VLAN.

4. ZAKOŃCZENIE

Współczesne heterogeniczne sieci korporacyjne stawiają przed administratorami wymagania zapewnienia jednolitego, bezpiecznego dostępu do zasobów sieci. Takim zasobem mogą być także dynamicznie zmieniające się grupy użytkowników, których członkowie wymieniają poufne, niedostępne dla innych pracowników informacje. Podstawowe wymaganie dotyczy zawsze gwarantowanego poziomu zabezpieczenia oraz możliwości śledzenia dynamicznego procesu dołączania i odłączania się użytkowników. Zastosowanie w sieciach VLAN z autoryzacją dostępu infrastruktury klucza publicznego umożliwia stworzenie bezpiecznych i efektywnych mechanizmów kontroli dostępu. Prezentowane rozwiązanie składa się z dwóch etapów, z których pierwszy jest weryfikacją tożsamości

klienta w oparciu certyfikat klucza publicznego zgodny ze standardem X.509. Po sprawdzeniu tożsamości algorytm bezpiecznie dystrybuuje aktualne symetryczne klucze sesji w celu utworzenia lub dostępu do bezpiecznego kanału wymiany informacji. Drugi etap to weryfikacja praw dostępu do określonego zasobu, którą w prezentowanym algorytmie sieci VLAN oparto na certyfikatach atrybutu zawierających listę udostępnionych wybranemu użytkownikowi zasobów. Po sprawdzeniu ważności certyfikatu atrybutu wybrany punkt dostępu (np. serwer sieciowy) umożliwia bezpieczne przesyłanie informacji. Przedstawiony w artykule algorytm może się stać podstawą do budowy niezależnych od platformy sprzętowej i programowej rozproszonych systemów autoryzacji dostępu do zasobów i usług w sieciach korporacyjnych.

BIBLIOGRAFIA

- [1] Górski G., Woźniak J.: *Self-defining virtual LANs (AUTO-VLANs) with access authorisation*. W: Proceedings of the International Scientific NATO PfP/PWP Conference Security and Protection of Information CATE 2001, May 9-11 2001, Brno, Czech Republic., t.I, s.22-28.
- [2] Górski G., Woźniak J.: *Dynamic virtual LANs with access authorisation..* W: Proceedings of IX Regional Conference On Military Communication and Information Systems 2000, October 4-6 2000, Zegrze, Poland, t.III s.89-93.
- [3] Górski G., Woźniak J.: *Metody zarządzania zasobami w nowoczesnych sieciach heterogenicznych*. W: Proceedings of Krajowe Sympozjum Telekomunikacji '99, September 8-10 1999 Bydgoszcz, Poland, t.C s.304-310.
- [4] Adams C., Lloyd S.: *Understanding Public-Key Infrastructure*, Macmillan Technical Publishing, 2001.
- [5] Farrell S., Housley R.: *An Internet Attribute Certificate Profile for Authorisation*. Network Working Group RFC 3281, April 2002
- [6] ITU-T Recommendation X.509 (03/2000) Series X: Data Networks and Open System Communications: *Information technology – Open systems interconnection – The Directory: Public-key and attribute certificate frameworks*.

SECURE VLAN NETWORKS WITH ACCESS AUTHORISATION BASED ON ATTRIBUTE CERTIFICATES

Summary

The paper presents the dynamic virtual LAN algorithm with access authorisation to different network resources. Authorisation process is a procedure which verifies a user identity and service requests. The access control process uses both public key certificate and attribute certificate precisely defined by X.509 ITU standard. The algorithm is a new solution which can be a part of distributed access control systems implemented in corporate network environment.

Henryk Krawczyk, Michał Wielgus

**Katedra Architektury Systemów Komputerowych
Politechnika Gdańska**

PAKIET OCENY BEZPIECZEŃSTWA SYSTEMÓW INFORMACYJNYCH¹

Streszczenie

Opracowano koncepcję oceny bezpieczeństwa w skali 6-cio poziomowej umożliwiającej lokalizację mechanizmów zabezpieczeń w różnych warstwach architektury systemu, a także różnych procedur reagowania w przypadku naruszenia bezpieczeństwa. Przedstawiono założenia logiki rozmytej oraz diagram przejść stanów przydatne do badania reakcji systemu na tzw. testy penetracyjne. Podano architekturę symulatora środowiskowego do oceny różnych polityk bezpieczeństwa i różnych mechanizmów zabezpieczających.

1. WSTĘP

Każdy system przechowujący, przetwarzający i/lub przesyłający informację pod różnymi postaciami, przy użyciu dowolnych środków stanowi w pojęciu autorów system informacyjny (w skrócie SI). Celem pracy jest automatyzacja procesu oceny bezpieczeństwa tak rozumianego systemu, co jest możliwe przy wykorzystaniu właściwego pakietu oceny bezpieczeństwa. W praktyce istnieje wiele takich pakietów wykorzystujących różne modele bezpieczeństwa systemów, jak i różne metody analizy bezpieczeństwa (TISM). Na ogół są to jednak specjalizowane narzędzia umożliwiające ocenę jednego z aspektów bezpieczeństwa (IDS snort, tripwire, DomiLock; snifery sniffit, hunt; skanery nmap, hping, saint; exploity). Praca zawiera próbę uogólnionego podejścia, w którym główne aspekty bezpieczeństwa będą rozpatrywane łącznie. Umożliwia to uwzględnienie zadanej polityki bezpieczeństwa, rozpatrzenie reprezentatywnych przypadków użycia oraz szacowanie możliwości zwiększenia bezpieczeństwa poprzez wykorzystanie wzorcowych mechanizmów zabezpieczających.

Koncepcja działania pakietu oceny bezpieczeństwa została przedstawiona w Punkcie 2. Opracowanemu modelowi bezpieczeństwa systemów informacyjnych poświęcono Punkt 3. Architektura pakietu jest zawarta w Punkcie 4. Na zakończenie przedstawiamy zalety pakietu i wnioski jego analizy.

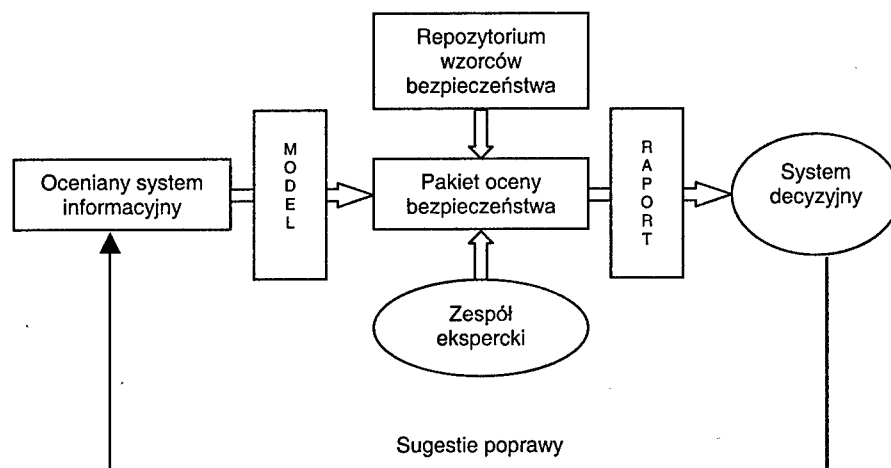
¹ Opracowano w ramach grantu KBN Nr 4T11C 005 25

2. KONCEPCJA

Pakiet oceny bezpieczeństwa systemów informacyjnych ma za zadanie analizę bezpieczeństwa i oceny ryzyka konkretnego przypadku, modelowanie jego zachowania się, tworzenie wzorców zachowań, a w ostateczności podanie szacunkowej oceny bezpieczeństwa (SOB). Przyjmuje się a priori założenie, że opracowany model (wzorzec) systemu bezpieczeństwa zawiera najlepsze możliwe rozwiązania ochronne, które wyrażają się poprzez wykorzystanie odpowiednich mechanizmów zabezpieczających. Analiza systemu uwzględnia zadaną politykę bezpieczeństwa oraz opiera się na testach penetracyjnych (potwierdzenie/zaprzeczenie istnienia takiego mechanizmu). Podejmowane akcje sprawdzające (działania zmierzające do ustalenia istnienia mechanizmów ochronnych) prowadzą do konkretnej reakcji systemu (odpowiedź na dane zapytanie). Po tym następuje interpretacja generowanych odpowiedzi, porównanie zgodności z odpowiedzią wzorcową oraz zapisanie otrzymanych wyników do bazy wiedzy, które w końcu pozwolą na modelowanie wzorca (elementy logiki rozmytej, diagram przejść stanów), opisanie trendów zachowania systemu (statystyka) oraz wyciągnięcia wniosków o możliwościach systemu. Na tej bazie możliwa jest szacunkowa ocena bezpieczeństwa na jednym z 6 poziomów bezpieczeństwa [1]:

- Poziom 1 – poddane analizie funkcjonalnej
- Poziom 2 – poddane analizie strukturalnej
- Poziom 3 – poddane systematycznej analizie i kontroli
- Poziom 4 – poddane systematycznej konstrukcji, analizie i inspekcji
- Poziom 5 – poddane półformalnej konstrukcji i analizie
- Poziom 6 – poddane półformalnej kontroli konstrukcji i analizie

Konieczna wydaje się wielopłaszczyznowa analiza bezpieczeństwa dotycząca architektury trójwarstwowej (infrastruktura sieciowa – warstwa pośrednicząca – aplikacje), której celem jest wybór wzorcowych mechanizmów zabezpieczeń, zapewniających wymagany poziom bezpieczeństwa. Stanowi to proces ciągły, który pozwala na modelowanie systemu informacyjnego w sposób zależny od wymagań użytkownika (rys.1). Zespół ekspercki jest odpowiedzialny za przypisanie ocen mechanizmom bezpieczeństwa w odniesieniu do wzorców bezpieczeństwa. Jest niezbędny, jeżeli chodzi o ocenę bezpieczeństwa z uwagi na fakt różnicowania mechanizmów zabezpieczających. Utworzone raporty zawierają szczegółowe informacje o środkach bezpieczeństwa wbudowanych w system, ocenę ich przydatności i skuteczności, wskazują na luki w zabezpieczeniach, podają zalecenia, co do poprawy przyjętych rozwiązań. System decyzyjny określa poziom bezpieczeństwa badanego systemu na podstawie zredagowanych raportów i pełni rolę nadrzędną w stosunku do zespołu eksperckiego. Takie rozwiązanie pozwala na samokontrolę działania oraz podnosi wiarygodność oceny. Poddając wymagania takiej analizie możemy rozbudowywać model bezpieczeństwa, co z kolei pozwala na ulepszenie mechanizmów zabezpieczeń tworząc nowy, ulepszony, bardziej bezpieczny system. Wyznaczone na etapie modelowania parametry zależne od bezpieczeństwa po przełożeniu na metryki pozwolą na ogólną ocenę bezpieczeństwa systemu. Zadania te realizuje przygotowany pakiet oceny bezpieczeństwa. Przewidywany zakres wykorzystania pakietu dotyczy tworzenia polityk bezpieczeństwa poszczególnych obszarów w zakresie ochrony fizycznej, bezpieczeństwa osobowego, bezpieczeństwa informacji, ciągłości działania oraz bezpieczeństwa SI.



Rys.1. Proces modelowania i oceny bezpieczeństwa systemu informacyjnego

3. MODEL BEZPIECZEŃSTWA SI

Model bezpieczeństwa systemów informacyjnych powinien uwzględniać takie czynniki jak niezależność od środowiska, ukierunkowanie na indywidualne cechy systemu, możliwość oceny systemu na każdym etapie cyklu życia. Przyjmuje się koncepcję matrycowego modelu bezpieczeństwa, który zakłada spełnienie określonych wymagań (zbieranie informacji o systemie w oparciu o kwestionariusze bezpieczeństwa, tworzenie raportów bezpieczeństwa, analiza danych zawartych w raportach) przed przystąpieniem do definiowania systemu informacyjnego w postaci szablonów bezpieczeństwa. Celem wymienionych działań jest wydobycie istotnych informacji z punktu widzenia bezpieczeństwa systemu, dokonanie wstępnej selekcji i określenie szacunkowych wartości poszczególnych elementów systemu [2].

Matrycowy model bezpieczeństwa ma dwójakie zastosowanie. Po pierwsze stanowi odwzorowanie bezpieczeństwa systemu informacyjnego w matematyczny model na danym poziomie bezpieczeństwa zgodnie z obowiązującymi kryteriami bezpieczeństwa (CC, ITSEC). Po drugie stanowi zestawienie parametrów systemu zapewniających określony poziom bezpieczeństwa w odniesieniu do wybranej cechy systemu. Definiowanie systemu informacyjnego w postaci modeli matrycowych może się odbywać w płaszczyźnie poziomej (wewnętrznej), pionowej i w głąb (zewnętrznej), tworząc kolejne warstwy wybranych cech na jednym z 6 poziomów bezpieczeństwa. W ten sposób zostaje utworzona wielowymiarowa macierz systemu. Jej wartościowanie dokonywane jest przez zespół ekspercki na podstawie raportów bezpieczeństwa opracowanych z kwestionariuszy bezpieczeństwa oraz wyników niezależnych testów penetracyjnych.

Szacowanie bezpieczeństwa systemu w oparciu o uzupełnione matryce sprowadza się do wyznaczenia wartości minimalnych otrzymanych wyników (tabela 3.1). Dane zestawione w matrycach bezpieczeństwa muszą jeszcze zostać przetworzone w sposób metodyczny dostarczając miarodajnej oceny bezpieczeństwa systemu. Jest to możliwe dzięki zastosowaniu łańcuchowej metody szacowania bezpieczeństwa. Posługując się tymi danymi możemy dokonać oceny bezpieczeństwa (3.1)

Tablica 3.1

Matrycowy model bezpieczeństwa

			Ocena ₁₁	Ocena ₁₂	...	Ocena _{1z}
		Ocena _{2z}
	Ocena ₁₁₂	Ocena ₁₂₂	...	Ocena _{1y2}
Ocena ₁₁₁	Ocena ₁₂₁	...	Ocena _{1y1}	Ocena _{2y2}	...	Ocena _{xy2}
Ocena ₂₁₁	Ocena ₂₂₁	...	Ocena _{2y1}
...	Ocena _{xy2}
Ocena _{x11}	Ocena _{x21}	...	Ocena _{xy1}

Ocena_{xyz} – ocena bezpieczeństwa danej cechy (x) na poziomie (y) w warstwie (z)

$$SOB = \min \{ Ocena_{xyz} \} \quad (3.1)$$

gdzie: *Ocena* – oznacza szacunkową ocenę bezpieczeństwa z kwestionariuszy,
x, y, z – indeksy macierzy odpowiednio *x* – badana cecha systemu (el. wewn.), *y* – poziom bezpieczeństwa, *z* – warstwa bezpieczeństwa (el. zewn.),
SOB – oznacza szacunkową ocenę bezpieczeństwa systemu,

W każdym przypadku należy wyznaczyć wartości minimalne przypisane na etapie wstępnej selekcji. Określona w ten sposób ocena pozwala na jednoznaczne określenie stopnia ochrony informacji, odslaniając potencjalne luki w systemie zabezpieczeń zgodnie z zasadą, że wytrzymałość łańcucha mierzy się trwałością jego najsłabszego ogniwa.

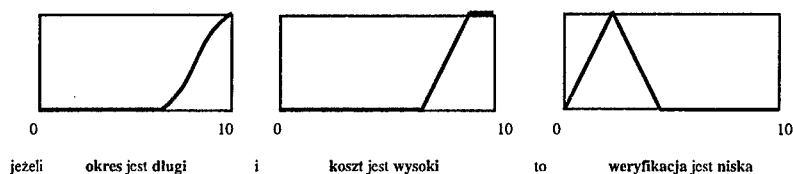
Tablica 3.2

Model bezpieczeństwa z ważoną powierzchnią bezpieczeństwa

2Ocena ₁₁₁	Ocena ₁₂₁ /5	...	Ocena _{1y1} /30
4Ocena ₂₁₁	Ocena ₂₂₁ /10	...	Ocena _{2y1} /40
...
20Ocena _{x11}	Ocena _{x21} /2	...	Ocena _{xy1} /50

N*Ocena_{xyz} – N-krotna weryfikacja oceny w danej warstwie macierzy

Modele rozmyte wprowadzają zmienne wagowe w celu rozróżnienia stopnia dokonywania powtórnej oceny (tabela 3.2 – liczby przy ocenach w pierwszej warstwie). Punktem wyjściowym jest uzasadniona gradacja przeprowadzania weryfikacji oceny bezpieczeństwa w zależności od czasu (okresu zachowania bezpieczeństwa) oraz kosztów bezpieczeństwa (nakładów poniesionych na ochronę informacji). Im dłuższy okres zachowania bezpieczeństwa, im większe koszty poniesione na ochronę informacji tym mniejsza częstotliwość weryfikacji oceny i odwrotnie, im krótszy okres zachowania bezpieczeństwa, im mniejsze koszty poniesione na ochronę tym wyższa częstotliwość weryfikacji (rys.2). Skala na rysunku została dobrana arbitralnie [3].



Rys.2. Przykładowe funkcje przynależności i reguły rozmyte

Takie postępowanie formuje powierzchnię ważonego bezpieczeństwa oznaczonej kolorem szarym (tabela 3.2 oraz tabela 3.3). Ustalenie progów bezpieczeństwa i przypisanie ich w miejsce wag macierzy umożliwia zastosowanie logiki rozmytej do weryfikacji oceny bezpieczeństwa (tabela 3.3).

Tablica 3.3

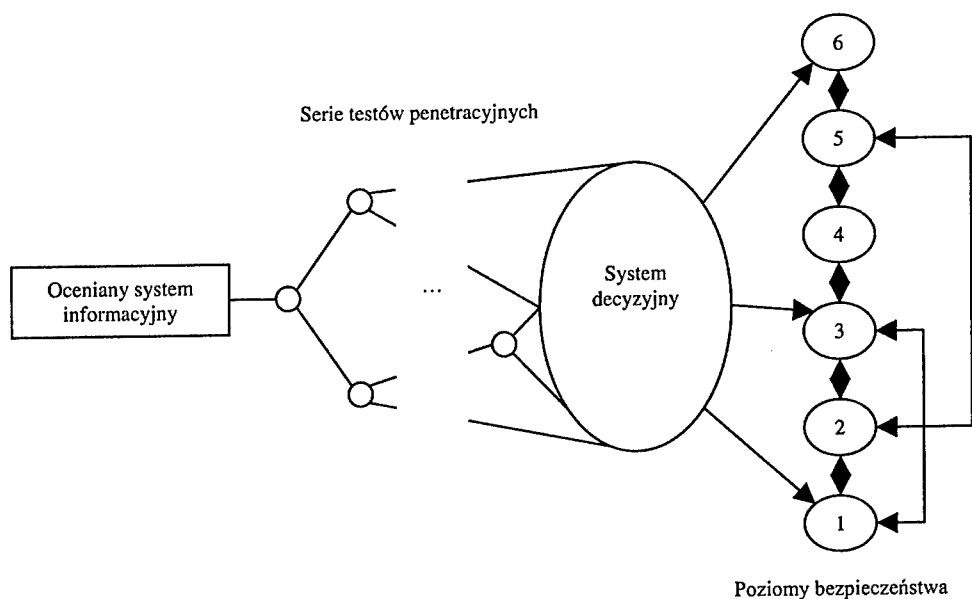
Rozmyty model bezpieczeństwa z ważoną powierzchnią bezpieczeństwa

DWO ₁₁₁	SWO ₁₂₁	...	MWO _{1y1}
WWO ₂₁₁	MWO ₂₂₁	...	NWO _{2y1}
...
BWVO _{x11}	DWO _{x21}	...	BNWO _{xy1}

WO_{xyz} – weryfikacja oceny bezpieczeństwa w danej warstwie macierzy
(odpowiednio: BW – bardzo wysoka, W – wysoka, D – duża,
S – średnia, M – mała, N – niska, BN – bardzo niska)

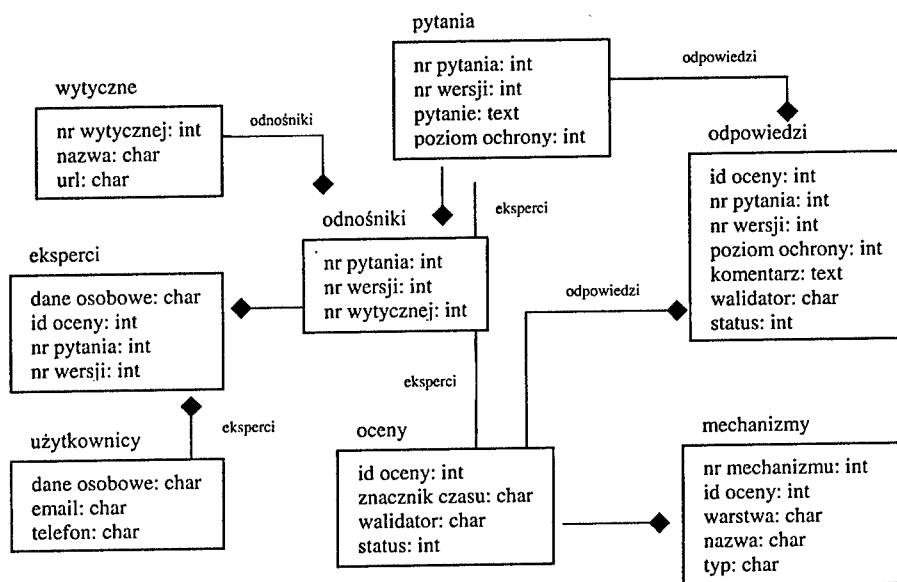
Podstawową słabością modeli opartych na systemach logiki rozmytej jest ich podatność na ujawnienie konturu, czyli ustalonego zasięgu powierzchni bezpieczeństwa, wewnątrz której weryfikacja ocen jest znikoma ze względu na wysoki poziom zastosowanych środków. Możliwe jest takie sterowanie ocenami, że żadna z nich nie będzie podlegała weryfikacji. Wykrycie spreparowanych ocen jest niezwykle trudne i oznacza dalsze traktowanie bezpieczeństwa w sposób, który zakłada brak okresu zachowania bezpieczeństwa, a odzyskanie stanu bezpieczeństwa możliwe jest dopiero po ustalonym upływie czasu. Jedną z metod zapobiegających tego typu działaniom jest przypadkowe zakłócanie porządku polegające na modulacji zasięgu powierzchni bezpieczeństwa z dowolną częstotliwością.

Testy penetracyjne systemu umożliwiają badanie reakcji systemu na działania celowe oraz niezamierzone. Pozwalają na modelowanie bezpieczeństwa systemu oraz zastosowanie w procesie oceny opinii zespołu eksperckiego. Wykorzystuje się tutaj diagram przejść stanów (rys.3). Rozpatrzmy przykład scenariusza testowego: zadaniem aplikacji jest szyfrowanie połączenia pomiędzy serwerem i klientem w oparciu o mechanizm kluczy szyfrujących. Na skutek błędnego zarządzania kluczami nie jest możliwe ustanowienie bezpiecznego połączenia. Sposobem na przetestowanie takiego przypadku jest czasowe wyłączenie mechanizmu zarządzania kluczami dla celów symulacji. Oceniany system poddawany jest serii takich testów. Wyniki testów poszczególnych etapów powodują wystawienie cząstkowej oceny, po ukończeniu wszystkich testów określana jest wartość SOB (3.1) i dokonuje się przypisanie systemu do określonego poziomu bezpieczeństwa. Jeżeli po powtórnych testowaniu systemu nastąpi zmiana oceny, jego poziom ulegnie zmianie.



Rys.3. Diagram przejść stanów w odpowiedzi na testy penetracyjne systemu

4. ARCHITEKTURA PAKIETU SOB



Rys.4. Schemat relacyjnej bazy danych systemu

Konstrukcja pakietu oceny bezpieczeństwa powstała na podstawie standardów OBI, ASSET [4]. Serce pakietu stanowi relacyjna baza danych (rys.4). Analiza bezpieczeństwa opiera się na szczegółowych kwestionariuszach bezpieczeństwa (tabela 4.1). Ich tworzenie i wypełnienie spoczywa na ekspertach w oparciu o wytyczne bezpieczeństwa [5]. Odpowiednie wytyczne mają oznaczenia literowo liczbowe (tabela 4.1 – FIPS 102, NIST SP 800-18). Dane do oceny pochodzą z testów penetracyjnych oraz narzędzi IDS. Oceny mogą przyjmować wartości z zakresu np. 1 – 7 (skala tabeli 3.3). Wskazane jest opatrzenie oceny szczegółowym komentarzem. Po wypełnieniu pól (tabela 4.1) następuje selekcja kwestionariuszy i utworzenie wielowymiarowego modelu bezpieczeństwa SI (tabela 3.1)

Funkcjonalność pakietu zawiera się w 5 podstawowych kategoriach:

1. Definicja i modyfikacja modelu bezpieczeństwa
2. Definicja i wykonywanie testów penetracyjnych
3. Selekcja wzorców bezpieczeństwa
4. Tworzenie i wykorzystanie reguł bezpieczeństwa
5. Wizualizacja poziomu bezpieczeństwa

Tablica 4.1

Przykładowy kwestionariusz bezpieczeństwa

Badana cecha systemu	Poziom ₁	Poziom ₂	Poziom ₃	...	Poziom ₆	Komentarz
Proces autoryzacji FIPS 102	Ocena ₁₁₁	Ocena ₁₂₁	Ocena ₁₃₁	...	Ocena ₁₆₁	Uwagi
4.1. Element krytyczny: Czy system uzyskał atest bezpieczeństwa potwierdzony niezależnie?	Ocena ₂₁₁	Ocena ₂₂₁	Ocena ₂₃₁	...	Ocena ₂₆₁	Uwagi
4.1.1 Czy proces certyfikacji został wykonany po znaczącej modyfikacji? NIST SP 800-18	Ocena ₃₁₁	Ocena ₃₂₁	Ocena ₃₃₁	...	Ocena ₃₆₁	Uwagi
4.1.2 Czy ocena ryzyka została przeprowadzona po modyfikacji? NIST SP 800-18
4.1.3 Czy reguły postępowania zostały utworzone i przedstawione użytkownikowi? NIST SP 800-18
4.1.4 Czy scenariusze zachowań zostały opracowane i wdrożone? NIST SP 800-18
4.1.5 Czy polityka bezpieczeństwa została uaktualniona i zweryfikowana? NIST SP 800-18
4.1.6 Czy mechanizmy bezpieczeństwa działają w zgodzie ze specyfikacją? NIST SP 800-18
4.1.7 Czy zastosowane mechanizmy są adekwatne do zagrożeń i wartości danych? NIST SP 800-18

4.1.8 Czy uzyskano autoryzację na współpracę z innymi systemami? NIST SP 800-18
4.2. Element krytyczny: Czy system działa w oparciu o rozwiązania tymczasowe zgodne ze specyfikacją?
4.2.1 Czy podjęto kroki zmierzające do poprawy deficytowego stanu systemu? NIST SP 800-18	Ocena _{x11}	Ocena _{x21}	Ocena _{x31}	...	Ocena _{x61}	Uwagi

Ocena_{xyz} – ocena bezpieczeństwa danej cechy (x) na poziomie (y) w warstwie (z)

5. ZAKOŃCZENIE

Etap budowania matrycowych modeli bezpieczeństwa systemów informacyjnych jest najbardziej czasochłonną częścią procesu szacowania bezpieczeństwa. Jest przy tym składową najbardziej newralgiczną. Nie istnieje jedna reguła mająca zastosowanie do analizy każdego systemu. Decydującą rolę odgrywa wrażliwość przesyłanej, przetwarzanej i przechowywanej informacji z punktu widzenia użytkownika systemu. Sama metoda łańcuchowa jest nieskomplikowana i znajduje zastosowanie z przypadku systemów ściśle informacyjnych, jak również w środowiskach hybrydowych. Zaletą proponowanego podejścia jest indywidualizacja modeli w odniesieniu do rozpatrywanego systemu.

BIBLIOGRAFIA

- [1] *Common Criteria 2 An Introduction*, W: Syntegra on behalf of the Common Criteria Project Sponsoring Organizations, May 1998
- [2] Krawczyk H., Wielgus M.: *Szacowanie bezpieczeństwa systemów informacyjnych*, W: Techniki Komputerowe, Rok XXXVII, Nr 1, 2002, Biuletyn Informacyjny, Instytut Maszyn Matematycznych, Warszawa 2002
- [3] Krawczyk H., Wielgus M.: *Modele bezpieczeństwa aplikacji rozproszonych*, W: Studia Informatica, Volume 23, Number 2B (49), Wydawnictwo Politechniki Śląskiej, Gliwice 2002
- [4] McLarnon M., Swanson M.: *Automated Security Self-Evaluation Tool Technical Documentation*, National Institute of Standards and Technology, Technology Administration, U.S. Department of Commerce, January 31, 2003
- [5] Swanson M.: *Security Self-Assessment Guide for Information Technology Systems*, National Institute of Standards and Technology, Technology Administration, U.S. Department of Commerce, November 2001

INFORMATION SYSTEMS SECURITY ASSESSMENT PACKAGE

Summary

Idea of security evaluation for information processing systems is suggested. Six security levels are taken into account. Fuzzy logic assumptions and state transition diagrams are presented and their suitability for penetration tests are discussed. Security assessment package architecture is given in details. Selection of the suitable security policy and secure mechanisms for achieving the required level of system security are given. Formal methods for consistency assurance and system modeling are also met in particular.

Henryk Malak

Department of Physics, University of Maryland Baltimore County, MD 21043

BIOWARFARE AGENT DETECTION WITH QUANTUM ENTANGLEMNT OF PHOTONS

Abstract

Quantum entanglement of photons phenomenon allows for highly sensitive, noiseless, long distance (up to 100 km) and fast spectral detection with atomic and molecular spectral resolutions. Therefore, quantum spectral technique is superior to other spectral techniques for biological and chemical weapons agents sensing. Our research and development is focused on the following quantum imaging effects: (1) Quantum imaging with entangled photon pairs, based on the working principle of our early "ghost" image experiment [Phys. Rev. Lett., Vol. 74, 3600 (1995); Phys. Rev. A, Rapid Comm., Vol. 52, R3429 (1995)]. (2) Entangled two-photon microscopy, based on our recent demonstration of quantum lithography experiment [Phys. Rev. Lett., Vol. 87, 013602(R) (2001)]. Due to the use of entangled photons quantum effects, the developed "quantum spectral" tools offer homeland security relevant applications..

1. INTRODUCTION

Novel concepts and techniques for point and/or remote spectral detection of BWA bio-aerosols by utilizing the nonlocal property of entangled two-photon states with BWA specific capture technology and advanced spectral algorithms is here presented. We believe that implementing these techniques into the current BWA trigger spectral sensor will lead to the developed sensor meet the goals of current needs of homeland security. A detail description of the technology of the BWA sensor is presented below.

2. QUANTUM SPECTROMETER FOR SPECTRAL SENSING OF BIO-AEROSOLS

By utilizing the nonlocal property of entangled two-photon state of spontaneous downconversion (SPDC), a novel remote quantum spectrometer for characterizing the spectral function of remote optical elements can be realized. The signal idler photon pair, either degenerate or nondegenerate is sent to distant locations such as a space station (signal) and a local ground laboratory (idler). On the space station, the registration time history of the signal, which passes through the spectral filtering object, is recorded. At the

same time, the registration time history of the idler, which passes through the scanning monochromator in the laboratory, is also recorded. The two individualized time history records are brought together by means of a classical communication channel. By analyzing the "coincidence" counting rate, the spectral function of the optical element is then obtained accordingly.

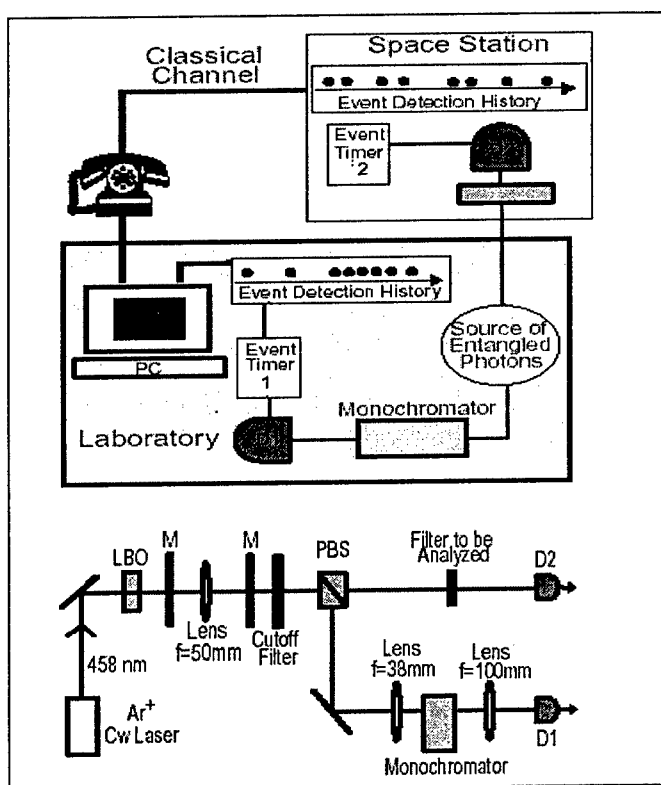


Figure 1. Scheme of a remote spectrometer and experimental setup

The schematic setup of the remote spectrometer is shown in **Figure 1**. Entangled photon pairs are generated in a type II nonlinear crystal through the spontaneous parametric down conversion process in a local laboratory. A polarizing beam splitter splits the orthogonal polarized photon pair, signal and idler. The signal photon is sent to a remote location, in which there is the optical element to be analyzed (e.g. a spectral filter). The idler photon passes through a monochromator in the laboratory. The signal and the idler are detected by photon counting detectors D_1 and D_2 , respectively. Each detector is connected to an event timer, an electronic device that records the registration time history at which a "click" detection event on the detector has occurred. The registration time history of detector D_1 located on the space station is sent back to the laboratory through a classical communication channel (telephone, Internet etc.). The two individual registration time histories are analyzed to achieve maximum "coincidences" by shifting the time bases of the

two. The remote spectrometer is now properly set. The spectral function of the remote spectral filter is obtained by measuring the rate of coincidence counts at each wavelength set by the local scanning monochromator.

While the remote operation of the spectrometer is fascinating, perhaps the most important feature of this device and technique is the enormous range of wavelengths that can be analyzed. This aspect of the remote spectrometer comes from the frequency correlation between the signal and the idler photons due to the phase matching condition of SPDC. According to quantum entanglement satisfying phase matching condition $\omega_p = \omega_s + \omega_i$ (where ω_j , j ($j = s, i, p$) are frequencies and of the signal (s), idler (i), and pump (p) respectively), if a pump laser at 400nm is used, a scanning monochromator working in the visible region (400nm - 700nm) will be able to remotely analyze a virtually infinitely large range of infrared wavelengths that starts from near infrared region. It is important to notice that the resolution achievable with the remote characterization will be determined mainly by the monochromator's inherent resolution. Thus, using a high-resolution monochromator in visible wavelengths will permit high-resolution calibration of spectral elements in near infrared and infrared wavelengths.

It also should be stressed that sensing, using quantum entanglement has an extremely low noise background. Theoretically, there should not be generated noise in this technique. Therefore, quantum entanglement is used successfully in cryptography where 100% interference events are expected. In our previous work we demonstrated noise free experiments related to quantum imaging and quantum lithography (Phys. Rev. Lett., Vol. 74, 3600 (1995); Phys. Rev. A, Rapid Comm., Vol. 52, R3429 (1995); Phys. Rev. Lett., Vol. 87, 013602(R) (2001)). The data shown in **Figure 2** is the actual counting rate with no subtractions or corrections of any type. This work was done with spontaneous parametric down conversion. We are also in the process preparing experiments using parametric amplifiers as light sources. This promises to produce bright, entangled beams, which would make image transfer more efficient at the price of more noise. However, it has been shown that one can still get very low noise images.

Figure 2 reports typical remote spectral measurements for band-pass filter centered at 885.6nm with bandwidth of 11nm . In the graph, we provided two scales of wavelengths as horizontal coordinates, referred to the signal and the idler wavelengths. These wavelengths can also be read as local "actually" measured wavelength (λ -idler) and "remote" indirectly measured wavelength (λ -signal).

The reported single detector counting rates of D_2 are slightly "tilted" at longer wavelengths. The tilting slope is mainly determined by the coupling efficiency of the monochromator. To account for this, we normalized the coincidence counts accordingly (see each of the figure captions for details). It is clear from the experimental data that the remote quantum spectral measurements agree with the classical spectral transmissivity calibration curves and with the theoretical predictions.

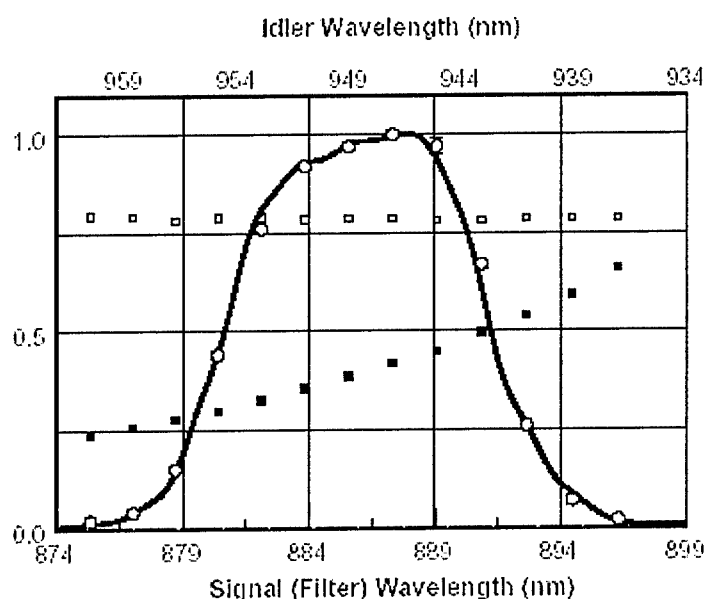


Figure 2. Experimental characterization of an 11nm band-pass filter centered at 885 nm. The solid line represents a direct measurement of the transmissivity function of the 885 nm spectral filter; hollow squares are the single counts of detector D_1 ; filled squares are the single counts of detector D_2 . The circles are the normalized coincidence counts weighted by the single counts of detector D_2 .

The presented above remote quantum spectrometer methodology and experimental results indicated the possibility of applications to the spectral bio-aerosols sensing program. This program will benefit from several unique features of the remote quantum spectrometer. First, wide range of spectrum can be used for BWA sensing with a single line of laser excitation. It is known that BWA has a wide spectral signature in ultraviolet – visible (UV-VIS) – infrared (IR). Hence, the use of wide range spectrum will enhance the resolution power of spectral algorithms. Secondly, our quantum spectrometer is almost electronic/photon noise free. Therefore, we expect the developed trigger sensor will significantly reduce the false alarm rate. Third, we also anticipate atomic/molecular spectrum resolution of the BWA trigger sensor. Such high resolution is again needed for strengthen algorithms in the model development and in the reduction false rate. Fourth, the remote operation range of the developed sensor can be extended to kilometers. It has been experimentally demonstrated by other groups that entanglement of photons is preserved up to 100 km. Therefore, the sensor can be used either on the Ground and/or in the Space. The satellite(s) spectral sensing will fundamentally increase the possibility of satisfying the special requirements. Fifth, reporting spectral sensing in real-time is another important feature of the developed sensor. An absorption event in the signal beam is observed at the same time in the idler beam.

3. BIOINFORMATICS DESIGN AND ANALYSIS OF STATISTICAL SIGNAL PROCESSING

Detection and identification of biomolecules, biological agents and chemical composition have been a great concern in biological science, environmental, industrial and defense applications. Unlike detection of rigid targets such as tanks, trucks and buildings which can be based on target spatial shapes and sizes, these targets do not have fixed sizes. Instead, they have unique spectral characteristics and are active in ultraviolet – visible (UV-VIS) and infrared (IR) ranges and cannot be identified by visual inspection. This presents a difficult and challenging problem since many existing traditional spatial-based image analysis techniques are generally not applicable. In biological and chemical detection, UV-VIS and IR sensing generally provides a valuable method for the detection and identification of biomolecules, biological agents and chemical composition. A chemical sensor or biosensor consists of an optical detector and a signal processor that operate in series to detect the presence of their spectral signatures. The detector measures the spectral signatures of interest and background scene, while the signal processor discriminates between these target signatures and the background signatures. Therefore, its detection performance is dependent upon the design of optical detection system and its statistical properties and the detection algorithms employed by the signal processor. In order to achieve this goal, we need to develop and design spectral-based signal processing algorithms for signature detection and identification. Hyperspectral imaging provides an effective means to meet such needs. Therefore, one of primary tasks of hyperspectral imaging is to design and develop spectral-based image processing (non-literal) techniques that can characterize and capture spectral properties of targets rather than their spatial properties for automatic detection and classification. This research investigates non-literal data exploitation for UV-VIS and IR spectrometry that cannot be generally solved by traditional spatial-based signal processing techniques. In particular, it will focus on exploration of new ideas, and design and development of non-literal signal processing analysis techniques for UV-VIS and IR spectrometry. Specifically associated with this process is the identification and interpretation of, subpixel spectral detection of biomolecules, biological agents, chemical composition, hazardous materials and toxic wastes in the environment.

4. DISCUSSION

The presented novel concepts and techniques for point and/or remote spectral detection of BWA bio-aerosols by utilizing the nonlocal property of entangled two-photon states with BWA specific capture technology and advanced spectral algorithms can be implemented to the current BWA trigger spectral sensor and will meet the goals of current needs of homeland security.

BIBLIOGRAPHY

- [1] Berzanskis A., et al., Phys. Rev. A 60, 1626 (1999)
- [2] Boto A.N., et al., Quantum Interferometric Optical Lithography: Exploiting Entanglement to Bear the Diffraction Limit, Physical Review Letters, 85, 2733 (2000)
- [3] D'Angelo M., M.V. Chekhova, and Y.H. Shih, Two-photon Diffraction and Quantum Lithography", Phys. Rev. Lett., 87, 013602 (2001).
- [4] D'Ariano G.M., Fortschr. Phys., 48, 579 (2000), G. M. D'Ariano, M. Rubin, M. F. Sacchi, and Y. Shih, Fortschr. Phys., 48, 599 (2000)
- [5] Pittman T.B., Y.H. Shih, D.V. Strekalov, and A.V. Sergienko, "Optical Imaging by Means of Two-Photon Entanglement", Phys. Rev. A, 52, R3429 (1995).

**DETEKCJA BRONI BIOLOGICZNO-CHEMICZNEJ Z UŻYCIEM
KWANTOWO ZWIĄZANYCH FOTONÓW****Streszczenie**

Zjawisko kwantowego efektu związanych fotonów pozwala na bardzo czułą, bezszumową, na długi dystans (do 100 km) oraz szybka spektralna detekcję, z atomowo-molekularną spektralną rozdzielczością. Dlatego, kwantowa spektralna technika ma przewagę nad innymi spektralnymi technikami w detekcji broni biologiczno-chemicznej. Nasze badania i rozwój tej techniki jest oparty na następujących kwantowych efektach: (1) Kwantowe obrazowanie związanych parami fotonów, w oparciu o pracujący efekt naszego wcześniej „duch” obrazowego doświadczenia [Phys. Rev. Lett., Vol. 74, 3600 (1995); Phys. Rev. A, Rapid Comm., Vol. 52, R3429 (1995)]. (2) Kwantowo związanych dwufotonowej mikroskopii, opartej o nasz niedawno zademonstrowany kwantowy litograficzny eksperyment [Phys. Rev. Lett., Vol. 87, 013602(R) (2001)]. Rozwój kwantowych efektów związanych fotonów, pozwoli na opracowanie „kwantowych spectralnych” technik ważnych w zastosowaniach dla bezpieczeństwa kraju.

Wojciech Neubauer, Józef Woźniak

Katedra Systemów Informacyjnych, Politechnika Gdańska

MECHANIZMY BEZPIECZEŃSTWA W BEZPRZEWODOWYCH SIECIACH 802.11

Streszczenie

Lokalne sieci bezprzewodowe (WLAN) to technologia, która w ostatnim czasie zyskuje coraz większą popularność. Sytuacji tej sprzyjają zarówno wygoda użytkowania, łatwość instalacji, jak i konkurencyjność cenowa coraz tańszych urządzeń do komunikacji radiowej. Z punktu widzenia użytkownika najważniejszymi zaletami sieci bezprzewodowych jest zapewnienie obsługi stacji mobilnych, a także umożliwienie bezproblemowego dostępu do zasobów sieci praktycznie z dowolnego miejsca obsługiwanego obszaru. W wielu przypadkach o wyborze rozwiązań bezprzewodowych zamiast przewodowych decyduje wysoka cena okablowania. Dotyczy to szczególnie obiektów zabytkowych oraz miejsc trudnodostępnych, gdzie prowadzenie kabli związane jest z poniesieniem dodatkowych kosztów. Z drugiej strony fakt, że podstawowym medium transmisyjnym w sieciach WLAN są fale radiowe, rodzi uzasadnione obawy o bezpieczeństwo tych sieci. Każdy standard WLAN podchodzi do tego problemu odmiennie. Niniejszy artykuł opisuje system zabezpieczeń, jaki przyjęty został przez organizację IEEE dla standardu 802.11 (często określanego mianem sieci Wi-Fi). W kolejnych rozdziałach wskazane zostały znane wady przyjętego rozwiązania i propozycje jakie rozważane są w nowo opracowywanym standardzie bezpieczeństwa sieci WLAN 802.11i.

1. WSTĘP

Obecnie procesem standaryzacji lokalnych sieci komputerowych mającym umożliwić kompatybilność urządzeń różnych producentów zajmują się dwie organizacje ETSI i IEEE. Prace prowadzone w Europie przez ETSI doprowadziły do powstania dwóch wersji standardu HiperLAN. Jednakże do chwili obecnej największą popularność zyskał standard IEEE 802.11 a w szczególności jego wersja 802.11b pracująca w paśmie 2.4 Ghz. W podstawowej wersji standard ten precyzuje budowę warstwy fizycznej oraz podwarstwy dostępu do medium. Niedługo po przyjęciu standardu okazało się, że jest on zbyt ubogi w stosunku do oczekiwań rynku. Coraz większa różnorodność transmitowanego ruchu w sieci zrodziła potrzebę jego rozwoju. Wychodząc na przeciw oczekiwaniom, w ramach IEEE powstało kilka grup rozwijających poszczególne funkcjonalności sieci WLAN, a wśród nich grupa 802.11e (opracowująca metody zarządzania jakością usług) czy 802.11f (zajmująca się tematyką przełączania – roamingiem). W niedługim czasie powstała także

grupa 802.11i, która rozpoczęła pracę nad zdefiniowaniem rozwiązań mających poprawić bezpieczeństwo sieci. Mimo że w podstawowej wersji standard 802.11 definiuje metodę szyfrowania transmitowanych danych opartą na protokole WEP, szybko okazało się, że mechanizm ten jest całkowicie nieskuteczny. Przede wszystkim pokazano, że protokół WEP jest zabezpieczeniem trywialnie łatwym do złamania. Ponadto standard nie nakłada obowiązku jego implementacji pozostawiając decyzję w rękach producentów sprzętu. W rezultacie większość dostępnych urządzeń nie miała zaimplementowanych żadnych mechanizmów ochrony.

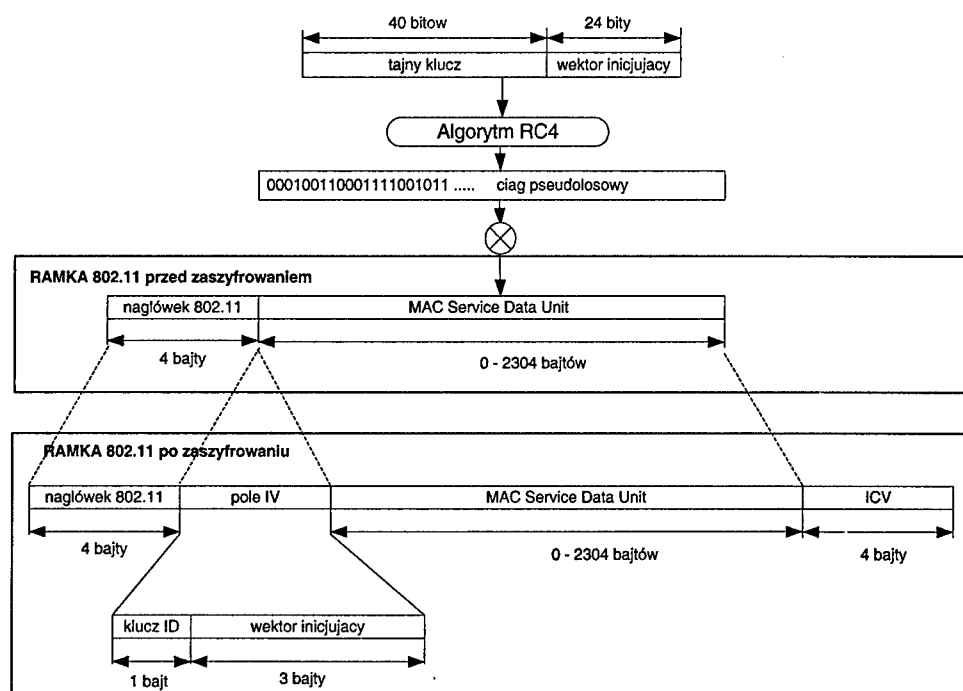
2. ZASADA DZIAŁANIA PROTOKOŁU WEP

Sieci bezprzewodowe bardzo trudno ograniczyć do ściśle zadanego obszaru. Fale radiowe z łatwością przenikają poza budynki, co powoduje, że każdy w bardzo łatwy sposób może próbować je odbierać. Skutkuje to większą podatnością sieci WLAN na różnego rodzaju ataki, co z kolei wymusza konieczność zastosowania dodatkowych mechanizmów zabezpieczających. Uwzględniając ten fakt, organizacja IEEE we wprowadzanym standardzie sieci bezprzewodowej 802.11 zdefiniowała protokół WEP (ang. *Wired Equivalent Protocol*) mający zapewnić bezpieczeństwo transmisji w warstwie łącza danych. W założeniach miał on zapewnić taki sam poziom bezpieczeństwa, jaki w naturalny sposób zapewniony jest przez sieć przewodowa, czyli ograniczonemu fizycznemu dostępowi do medium.

Zgodnie z założeniami protokół WEP miał zapewniać:

- poufność transmitowanych danych – sieć powinna być odporna na ataki pasywne polegające na podsłuchiwaniu transmitowanych ramek. Ponieważ bardzo trudno jest zapobiec nieautoryzowanemu odbiorowi fal radiowych, systemy zabezpieczeń powinny uniemożliwiać zrozumienie podsłuchanych danych. Protokół WEP wykorzystuje w tym celu stworzony przez Rona Rivesta w 1987 algorytm szyfrujący RC4.
- kontrolę dostępu – do sieci powinni mieć dostęp tylko autoryzowani użytkownicy. Sieć bezprzewodowa powinna uniemożliwić nieautoryzowanego korzystania ze swoich zasobów.
- integralność danych – sieć powinna zabezpieczać przed możliwością modyfikowania, a tym samym fałszowania transmitowanych wiadomości. Do ochrony integralności danych protokół WEP wykorzystuje powszechnie stosowaną sumę kontrolną opartą na kodzie cyklicznym CRC32.

Działanie protokołu WEP oparte jest na algorytmie RC4, który służy jako szyfrator strumieniowy. Dla każdego zadanego 64 bitowego klucza, algorytm ten zwraca pewien ustalony strumień bitów, który następnie zostaje wykorzystany jako ciąg szyfrujący. Dokładny schemat szyfrowania realizowany przez protokół WEP pokazano na rys. 1. Właściwe szyfrowanie polega na wykonaniu bitowej operacji XOR ciągu otrzymanego z wyjścia algorytmu RC4 z polem danych (MAC SDU) ramki przeznaczonej do wysłania. Dodatkowo w celu ochrony integralności transmitowanych danych wraz z ramką, zaszyfrowane zostaje pole ICV (*Integrity Check Value*) zawierające sumę kontrolną ramki. Wartość ta wyznaczana jest za pomocą algorytmu CRC32 obliczanego na danych zawartych w niezaszyfrowanej ramce.



Rys. 1. Zasada działania protokołu WEP oraz budowa ramek 802.11 przed i po zaszyfrowaniu

Najważniejszym aspektem bezpieczeństwa przy stosowaniu szyfrów strumieniowych jest odpowiedni schemat zmiany klucza po każdej transmisji, tak aby nie szyfrować dwóch wiadomości tym samym kluczem. W przypadku protokołu WEP 64-bitowy klucz używany do szyfrowania składany jest z dwóch części: 40 bitowego klucza tajnego oraz 24 bitowego wektora inicjującego (IV) przesyłanego jawnie wraz z każdą zaszyfrowaną ramką. Znajomość klucza tajnego jest warunkiem wystarczającym do szyfrowania i deszyfrowania ramek transmitowanych w warstwie łącza danych. Klucz ten jest zwykle współdzielony przez wszystkie autoryzowane stacje i punkty dostępowe sieci. 24-bitowy wektor inicjujący ma za zadanie zapewnić niepowtarzalność kluczy używanych przy transmisji kolejnych ramek. Aby ten cel zrealizować powinien zostać zaprojektowany algorytm wyboru kolejnych wektorów inicjujących, który zapewniałby niepowtarzalność klucza przed wykorzystaniem każdego dostępnego elementu z dostępnej przestrzeni 2^{24} liczb. Niestety standard nie definiuje żadnego tego typu algorytmu, co gorsza dopuszcza stosowanie tego samego wektora inicjującego dla kolejnych ramek. Dodatkowym niedociągnięciem standardu jest brak jakiegokolwiek mechanizmu dystrybucji kluczy czy tworzenia kluczy sesyjnych, co w praktyce oznacza konieczność ich ręcznego wpisywania w każdej stacji. Powoduje to, że wiele sieci korzysta z tych samych kluczy przez długie okresy czasu, rzędu miesięcy, co czyni je dość podatnymi na ataki. Niektórzy producenci próbując poprawić

bezpieczeństwo umożliwia definiowanie w urządzeniu tablicy 4 różnych tajnych kluczy, a informacja, który z nich wykorzystywany jest do szyfrowania, wysyłana jest w polu klucz ID ramki.

Zasada działania odbiornika/deszyfratora jest dokładnie odwrotna. Odbiornik wykorzystując znany mu tajny klucz oraz biorąc wektor inicjujący z odebranej ramki generuje za pomocą algorytmu RC4 pseudolosowy ciąg szyfrujący, który następnie wykorzystywany jest do operacji XOR z zaszyfrowanym polem ramki. W ten sposób otrzymuje się oryginalną treść wiadomości wraz z jej sumą kontrolną obliczoną przez nadajnik. Odbiornik porównuje tę sumę z sumą kontrolną obliczoną przez siebie i w przypadku gdy porównanie wypadnie pomyślnie ramka jest akceptowana. W wypadku przeciwnym ramka traktowana jest jako błędna.

3. SŁABE PUNKTY PROTOKOŁU WEP

Podczas projektowania za najsłabszy punkt protokołu WEP uważano niewielką długość klucza tajnego – efektywny 40 bitowy klucz jest dostatecznie krótki, aby próbować go złamać metodami siłowymi (ang. *Brute force attack*) i to bez konieczności użycia potężnych stacji obliczeniowych. W konsekwencji niektórzy producenci zaczęli implementować w swoich urządzeniach nie ujęty w standardzie protokół WEP2. Miał on zwiększyć poziom bezpieczeństwa sieci poprzez zastosowanie klucza 128 bitowego (100 bitów przeznaczonych było na klucz tajny i 24 bitów na wektor inicjujący). Niedługo po przyjęciu standardu pokazano, że podstawową wadą protokołu WEP nie jest zbyt krótki klucz tajny, lecz niewłaściwy sposób wykorzystania szyfratora strumieniowego. Udowodniono, że protokół ten nie jest w stanie spełnić postawionych założeń bez względu na długość zastosowanego klucza.

Podstawową wadą protokołu WEP jest to, że dopuszcza on możliwość szyfrowania dwóch kolejnych ramek za pomocą takiego samego klucza – używając tego samego wektora inicjującego. Oznaczając te ramki przez $R1$, $R2$ możemy zapisać:

$$R1 = P1 \oplus RC4(\text{wektor inicjujący, klucz})$$

$$R2 = P2 \oplus RC4(\text{wektor inicjujący, klucz})$$

gdzie:

$P1, P2$

$RC4(\text{wektor inicjujący, klucz})$

– ramki niezaszyfrowane

– pseudolosowy ciąg wygenerowany przez algorytm RC4 z użyciem klucza uzyskanego przez złożenie wektora inicjującego i współdzielonego klucza tajnego

Z powyższego wynika, że złożenie dwóch zaszyfrowanych ramek za pomocą operacji XOR powoduje, że wynik jest tożsamy z operacją XOR na dwóch niezaszyfrowanych ramkach:

$$R1 \oplus R2 = (P1 \oplus RC4(\text{wektor inicjujący, klucz})) \oplus (P2 \oplus RC4(\text{wektor inicjujący, klucz})) = P1 \oplus P2$$

Równanie to dowodzi, że dwie ramki zaszyfrowane tym samym kluczem ujawniają informację o niesionych wiadomościach. Do złamania zaszyfrowanej ramki R wystarczy znajomość treści jednej z wysłanych wiadomości $P1, P2$. W rzeczywistych warunkach jest to dość łatwe do spełnienia, wystarczy że atakujący wyśle do jednego z użytkowników sieci list o niewzbudzającej podejrzeń treści reklamowej, zna zawartość dynamicznych bibliotek przesyłanych przez serwery plików lub wykorzysta rozkład nagłówek i znane adresy w protokołach warstw wyższych. Nawet nie znając żadnej z wiadomości $P1, P2$ do ustalenia ich treści można wykorzystać nadmiar informacji językowej w tekście tj. rozkład częstości występowania poszczególnych liter alfabetu, słownik i kontekst wiadomości. Problem staje się jeszcze bardziej trywialny, gdy do dyspozycji są nie dwie, lecz więcej ramek zaszyfrowanych tym samym kluczem. Ich analiza umożliwia uzyskanie ciągu szyfrującego stosowanego dla konkretnego wektora inicjującego. W efekcie możliwe jest stworzenie całego słownika ciągów szyfrujących odpowiadających każdemu wektorowi inicjującemu, który pozwoli na deszyfrację każdej ramki w sieci. Dodatkowo, posiadając 24 bity klucza oraz ciąg szyfrujący, można próbować złamać sam klucz.

Przedstawione techniki niosą za sobą poważne konsekwencje dla bezpieczeństwa protokołu oraz dowodzą, że użycie tego samego wektora inicjującego do zaszyfrowania więcej niż jednej ramki czyni sieć 802.11 praktycznie otwartą na ataki z zewnątrz. Uwzględniając fakt, że długość wektora inicjującego wynosi 24 bity oraz zakładając, że stacja transmituje pakiety 1500 bajtowe z prędkością 5 Mbps, można pokazać, że czas po jakim zaczną następować kolizje IV (powtórzenia wektora inicjującego) jest mniejszy niż 10 godzin. W niektórych błędnych implementacjach czas ten jest jeszcze krótszy. Przykładowo losowanie wektora inicjującego dla kolejnych ramek powoduje, że prawdopodobieństwo kolizji po 5000 pakietów wynosi 50 %, natomiast po 11000 pakietów jest prawie pewne. W przypadku algorytmów zwiększających wartość wektora inicjującego o jeden po transmisji każdej ramki oraz jego zerowanie po każdej inicjacji interfejsu (restarcie urządzenia) powoduje że początkowe wartości z całej dostępnej przestrzeni liczb 24 bitowych są używane znacznie częściej niż wartości z końca tego przedziału.

Najważniejszym jednakże pozostaje fakt, że żadna implementacja mająca pozostać w zgodzie ze standardem nie jest w stanie wyeliminować kolizji wektorów inicjujących. Jest to jeszcze bardziej niebezpieczne w przypadku braku jakiegokolwiek mechanizmu kontroli i wymiany kluczy.

Oprócz zapewnienia tajności protokół WEP powinien gwarantować integralność transmisji. Jest to realizowanie poprzez obliczanie sumy kontrolnej CRC32 na niezaszyfrowanej wiadomości i włączeniu jej do zaszyfrowanej ramki. Ze względu na swoją liniowość CRC posiada następującą właściwość:

$$\text{CRC}(P1 \oplus P2) = \text{CRC}(P1) \oplus \text{CRC}(P2)$$

Umożliwia to zmienianie dowolnych bitów w zaszyfrowanej ramce w taki sposób, aby nie uszkodzić jej sumy kontrolnej. Zakładając, że przesyłana ramka R jest postaci:

$$R = \langle W, \text{CRC}(W) \rangle$$

gdzie :

- | | |
|-----------------|---------------------------------------------------|
| W | – zaszyfrowana wiadomość transmitowana w ramce |
| $\text{CRC}(W)$ | – suma kontrolna CRC obliczona dla wiadomości W |

natomiast fałszywa wiadomość W' jest utworzona z wiadomości W poprzez:

$$W' = W \oplus \Delta$$

widać, że aby zmienić wiadomość W na wiadomość W' , wystarczy w transmitowanej ramce P zmienić pole sumy kontrolnej przez poddanie jej operacji XOR z wartością $CRC(\Delta)$. Otrzymana w ten sposób zmieniona ramka będzie postaci:

$$R = \langle W \oplus \Delta, CRC(W) \oplus CRC(\Delta) \rangle$$

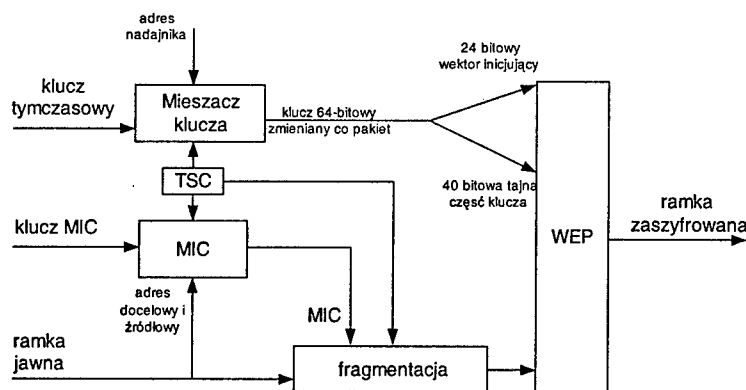
Oznacza to, że algorytm CRC nie jest w stanie efektywnie chronić integralności transmitowanych danych. Dodatkową dziurą protokołu WAP jaką stwarza wykorzystanie CRC jest jego niezależność od klucza szyfrującego. Faktycznie wystarczy znajomość jedynie strumienia bitów generowanego przez RC4 aby umożliwić atakującemu swobodną generację poprawnego strumienia ramek z właściwymi sumami kontrolnymi. Deszyfrator zgodny ze standardem 802.11 nie dysponuje żadnym mechanizmem umożliwiającym odrzucenie takiego ruchu.

4. PROTOKÓŁ TKIP

Z uwagi na naturę ujawnionych wad protokołu WEP jedynym słusznym rozwiązaniem było opracowanie całkowicie nowego protokołu, który byłby w stanie zagwarantować odpowiednio wysoki poziom bezpieczeństwa transmisji w sieci 802.11. Z drugiej strony liczba sprzedanych urządzeń, a za nią koszt ich wymiany, była na tyle duża, że powstała potrzeba opracowania rozwiązania łąającego WEP, możliwego do zastosowania poprzez aktualizację oprogramowania. Ponieważ starszy i wolniejszy sprzęt trzeba było zmusić do wykonywania dodatkowych obliczeń, podstawowym wymaganiem, które postawiono nowemu rozwiązaniu był jego niewielki nakład obliczeniowy. Tak powstała nakładka TKIP (ang. *Temporal Key Integrity Protocol*) owijająca istniejący protokół WEP.

Struktura funkcjonalna TKIP została przedstawiona na rysunku 2. W jej skład wchodzi:

- Mieszacz klucza, który dostarcza protokołowi WEP kombinacje IV i klucza służące do szyfrowania kolejnych ramek.
- TSC (ang. *TKIP Sequence Counter*) Kontrola kolejności wektorów inicjujących. Funkcja ta wymusza odrzucenie pakietów otrzymanych w złej kolejności
- MIC (ang. *Message Integrity Code*) nowa suma kontrolna zastępująca funkcjonalnie CRC obliczana nie tylko na podstawie wysyłanej treści, ale także tajnego klucza.
- 48 bitowy wektor inicjujący funkcjonujący analogicznie jak w protokole WEP, lecz jego przestrzeń możliwych liczb jest 2^{24} razy większa



Rys. 2. Zasada działania protokołu TKIP

Podstawową funkcją realizowaną przez TKIP jest uodpornienie systemu na kolizje wektora inicjującego. Osiągnięte to zostało poprzez zastosowanie dwufazowego mieszacza klucza. Jego zadaniem jest generacja klucza tajnego i wektora inicjującego, które są używane przez algorytm WEP do szyfrowania. Aby każda stacja w sieci 802.11 generowała inne klucze, mieszacz przekształca klucz tymczasowy w tak zwany klucz zmodyfikowany. Operacja ta, określana także jako faza pierwsza mieszacza, polega na połączeniu klucza tymczasowego z adresem nadajnika. Jest ona wykonywana raz w czasie obowiązywania klucza tymczasowego. W fazie drugiej klucz zmodyfikowany razem z sekwencyjnym licznikiem pakietów za pomocą prostego algorytmu szyfrującego przekształcany jest w wartość 128-bitową. Poprzez odpowiednie maskowanie wykluczone są wartości stanowiące tzw. słabe klucze algorytmu RC4 [4]. Otrzymany wynik rozdzielany jest na wektor inicjujący oraz klucz tajny, które służą algorytmowi WEP do zaszyfrowania kolejnego i tylko jednego pakietu.

Do zabezpieczenia transmisji przed próbami zmiany zawartości ramek lub nadawania ramek fałszywych użyto algorytmu MIC (wym, *Michael*). Posługując się 64-bitowym „kluczem MIC” dokonuje on obliczenia 64-bitowej sumy kontrolnej na niezaszyfrowanej ramce, a następnie przesyła ją wraz z ramką do stacji docelowej w celu umożliwienia weryfikacji jej autentyczności. Mimo że funkcja ta składa się jedynie z prostych operacji XOR i obracania bitów, stanowi ona główny narzut obliczeniowy dla stacji 802.11 pierwszej generacji. Jako kompromis pomiędzy skutecznością a szybkością działania nie stanowi ona także idealnego zabezpieczenia. Zakładany poziom jej bezpieczeństwa wynosi jedynie 20 bitów, co przy obliczanej 64 bitowej sumie kontrolnej nie jest wartością oszołamiającą. Oznacza to, że mogąc generować 2^{12} krótkich pakietów na sekundę, atakujący byłby w stanie stworzyć poprawny pakiet po 2^7 sekund (około dwóch minut). Analizy te wymusiły konieczność implementacji dodatkowych zabezpieczeń protokołu TKIP, które wzmocniły by jego odporność na fałszywe ramki. Jeżeli podczas odbioru stacja wykryje dwie ramki z błędnymi sumami kontrolnymi rozłącza się, wymienia klucze, czeka minutę i ponownie łączy się z siecią. Mimo że procedura ta znacząco może zaburzyć transmisję, umożliwia ono rozciągnięcie czasu w jakim możliwa byłby akceptacja fałszywego pakietu do około jednego roku.

Jak powiedziano na wstępie protokół TKIP jest rozwiązaniem mającym zapewnić obecnym sieciom 802.11 stosunkowo wysoki poziom bezpieczeństwa bez konieczności ponoszenia ogromnych wydatków na wymianę wszystkich urządzeń. Wprowadzone usprawnienia praktycznie wyeliminowały a przynajmniej teoretycznie zminimalizowały wszystkie słabe punkty dotychczasowego rozwiązania.

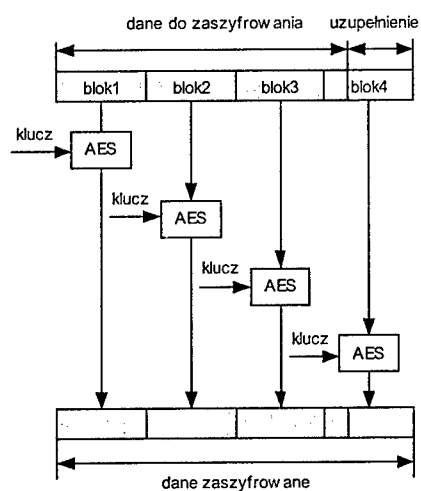
5. SZYFROWANIE AES

Protokół TKIP nie jest rozwiązaniem docelowym, które zapewniałoby wysoką tajność transmisji w sieciach 802.11, dlatego grupa standaryzująca opracowała zupełnie nowe rozwiązanie, oparte na nowym algorytmie szyfrującym AES (ang. *Advanced Encryption Standard*). Szyfr ten został przyjęty jako rządowy standard szyfrowania w Stanach Zjednoczonych zastępując algorytm DES z roku 1977. AES jest symetrycznym szyfrem blokowym pracującym na blokach danych o długości 16 bajtów. W przeciwieństwie do szyfru strumieniowego oznacza to, że AES do działania potrzebuje pełny blok danych, w związku z tym jeżeli kodowana wiadomość nie jest całkowitą wielokrotnością długości bloku musi ona zostać odpowiednio uzupełniona. Szyfrator AES może wykorzystywać klucze 128, 192 lub 256 bitowe, chociaż uważa się, że przy obecnie dostępnych mocach obliczeniowych klucz 128 bitowy jest wystarczająco bezpieczny (wystarczająco trudny do złamania). Zastosowanie szyfratora blokowego polega na wyborze odpowiedniego trybu pracy.

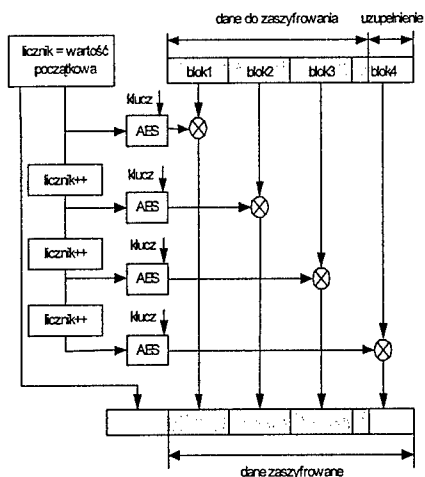
Rysunek 3 przedstawia trzy przykładowe tryby pracy szyfratora: ECB (ang. *Electronic Codebook mode*), CTR (ang. *Counter mode*), CBC (ang. *Cipher-Block Chaining mode*).

Pierwsza i zarazem najprostsza z przedstawionych konfiguracji (rysunek 3a) polega na podzieleniu wiadomości na bloki i poddanie każdego bloku niezależnej szyfracji. Ze względu na fakt że każdy blok szyfrowany jest tym samym kluczem ta sama treść bloku daje w wyniku taki sam ciąg szyfru. Powoduje to niepotrzebny wyciek informacji i osłabia skuteczność tej metody. Wady tej nie mają pozostałe konfiguracje. W konfiguracji CTR do szyfrowania wykorzystywany jest dodatkowy licznik. Na początku wartość licznika ustawiana jest na przyjętą wartość początkową, a następnie jest zwiększana o jeden dla każdego kolejnego bloku. Wartość licznika podawana jest na wejście szyfratora, który na podstawie tajnego klucza produkuje 128 bitowy ciąg pseudolosowy. Ciąg ten przy pomocy operacji XOR służy do zaszyfrowania jednego z bloków wiadomości. Początkowa wartość licznika dodawana jest w postaci jawnej do transmitowanej ramki, tak aby umożliwić drugiej stronie poprawną deszyfrację. Podobnie jak w przypadku WEP aby uniknąć drastycznego spadku poziomu bezpieczeństwa w tym trybie należy w implementacji wykluczyć powtórne użycie tego samego licznika. Trzeci zaprezentowany na Rysunku 3c tryb pracy szyfratora AES opiera się na wstępnym przekształceniu każdego bloku wiadomości z użyciem pewnej wartości pseudolosowej. Przekształcenie to jest prostą operacją XOR, która wykonywana jest dla pierwszego szyfrowanego bloku z losowym wektorem inicjującym, a następnie metodą iteracyjną z wynikiem szyfrowania poprzedniego bloku. Tak samo, jak w przypadku trybu CTR, w trybie CBC wartość wektora inicjującego dołączana jest w sposób jawny do transmitowanej ramki.

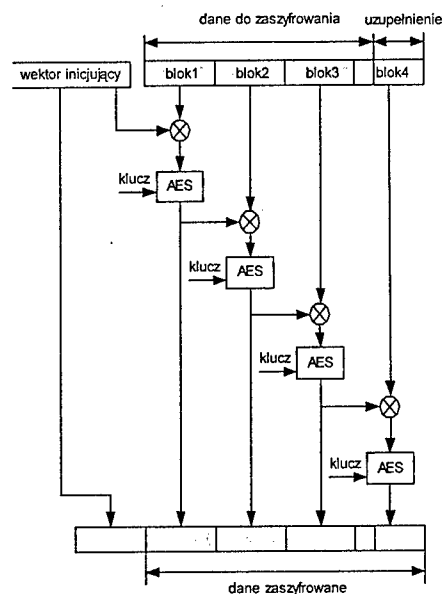
a) tryb ECB



b) tryb CTR

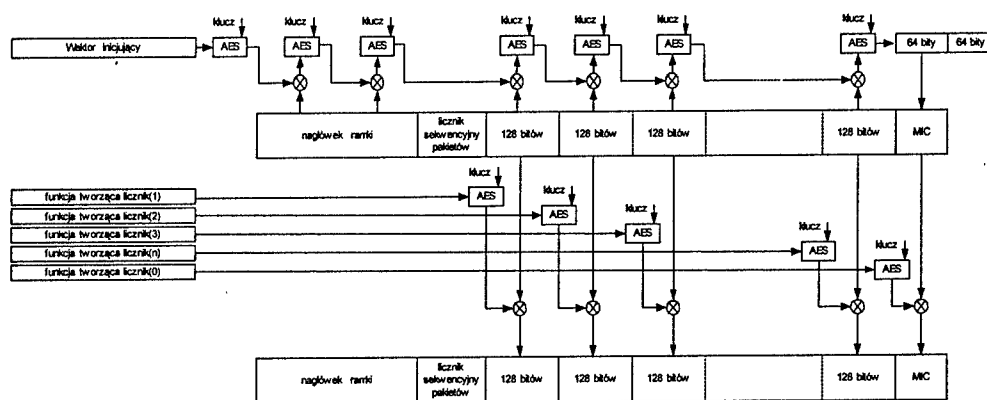


c) tryb CBC



Rys. 3. Trzy podstawowe tryby pracy szyfratora blokowego AES

Jedną z propozycji jaka jest rozważana przez grupę roboczą I organizacji IEEE jest protokół AES-CCM (ang. *Counter Mode with CBC-MAC*). Rozwiązanie to wykorzystuje dwa tryby pracy szyfratora AES: CBC-MAC i CTR. Pierwszy (CBC-MAC) chroni integralność transmisji służąc do wyznaczenia sumy kontrolnej MIC, natomiast drugi (CTR) chroni tajność danych i służy do szyfrowania pola danych ramki. Generalna zasada działania protokołu AES-CCM została przedstawiona na rysunku 4. Przedstawiony algorytm wykorzystuje ten sam klucz do operacji wyznaczania sumy kontrolnej oraz do właściwego szyfrowania.



Rys. 4. Zasada działania protokołu AES-CCM

Oprócz klucza algorytm CCM posiada 6 bajtowy sekwencyjny licznik pakietów. W oparciu o niego oraz w oparciu o adres źródłowy MAC, 16-bitowy licznik bloków i blok innych danych wyciąganych z nagłówka ramki tworzone są wartości początkowe wektora inicjującego i licznika, które potrzebne są w odpowiednich konfiguracjach pracy szyfratora. Samo szyfrowanie rozpoczyna się obliczeniem sumy kontrolnej MIC na adresie źródłowym, adresie docelowym, informacjach zawartych w polu QoS, polu długości ramki i polu danych. Następnie obliczona 128 bitowa suma kontrolna zostaje skrócona do 64 bitów i dołączana na końcu ramki. Powstała struktura zostaje zaszyfrowana przy użyciu trybu licznikowego szyfratora. Na koniec do ramki dodawany jest sekwencyjny licznik pakietów, który posłuży stronie odbiorczej do poprawnego odszyfrowania wiadomości.

6. PODSUMOWANIE

W opracowaniu przedstawione zostały podstawowe mechanizmy szyfrowania w warstwie łącza danych ujęte w podstawowym standardzie 802.11. Następnie pokazano, że mechanizmy te, oparte na protokole WEP, nie stanowią skutecznej ochrony sieci. Druga część artykułu zawierała opis wyników prac nad nowymi rozwiązaniami, które jak się obecnie sądzi nie posiadają żadnych z pierwotnych błędów. Wśród przedstawionych protokołów znalazły się nowe elementy standardu wchodzące w skład 802.11i, takie jak protokół TKIP czy AES-CCM. Ponieważ spodziewane jest ich rychłe pojawienie się na rynku jeszcze przed ogłoszeniem nowego standardu organizacja certyfikująca WiFi Alliance uruchomiła program certyfikacji urządzeń co do zgodności z przedstawionymi rozwiązaniami. Urządzenia które uzyskają tak zwany certyfikat WiFi Protected Access będą w pełni zgodne z przyszłym standardem.

LITERATURA:

- [1] Jesse Walker "Unsafe at any Key Size: An Analysis of the WEP Encapsulation", Tech. Rep. 03628E, IEEE 802.11 committee, March 2000.
- [2] Jesse Walker "802.11 Security Series Part II: The Temporal Key Integrity Protocol (TKIP)"
- [3] Rogaway P., M. Bellare, J. Black, T. Korvetz, "OCB Mode", April 1, 2001, <http://www.cs.ucdavis.edu/~rogaway/ocb/ocb.htm>
- [4] Fluhrer, Scott, Itsik Mantin, and Adi Shamir. "Weaknesses in the Key Scheduling Algorithm of RC4", Eighth Annual Workshop on Selected Areas in Cryptography, August 2001.
- [5] IEEE Std 802.11, Standards for Local and Metropolitan Area Networks: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.
- [6] Ferguson, N., "Michael: an Improved MIC for 802.11 WEP," IEEE 802.11 doc 02-020r0, January 17.
- [7] Stanley, D., "TV Sequencing Requirements Summary," IEEE 802.11 doc 02-006r2, January 18, 2002.
- [8] Housely, R., and D. Whiting, "Temporal Key Hash," IEEE 802.11 doc 01-550r1 October 31, 2001.

SECURITY MECHANISMS IN 802.11 WIRELESS NETWORKS

Abstract

Wireless Local Area Networks (WLAN) is a technology increasingly getting on demand with business and home clients. The fact is favored by the user convenience as well as the increasing competitiveness of still getting cheaper radio transmitters. In many cases it is high cost of laying the cable that inclines buyers to choose wireless solutions. It is particularly the case in historical buildings or places where it is extremely difficult to fix the cables. From the user perspective the most important advantage of WLANs is the support for mobile stations and the ability to get serviced from any place in the range of the network. On the other hand wireless networks are much more vulnerable in terms of transmission security. This is caused by the fact that radio waves used in communication

can easily permeate beyond a desired region of service. Various standards deal with the issue differently. This article focuses of security mechanisms worked up by IEEE organization in its 802.11 standard. Successive chapters cover flaws of the accepted solution and then describe new ideas which are considered for new security standard 802.11i.

Marek Niedostatkiwicz

Katedra Metrologii i Systemów Elektronicznych, Politechnika Gdańska

KONCEPCJA ROZPROSZONEGO SYSTEMU ZABEZPIECZAJĄCEGO POJAZD SAMOCHODOWY OPARTEGO NA DEDYKOWANYCH MAGISTRALACH CYFROWYCH

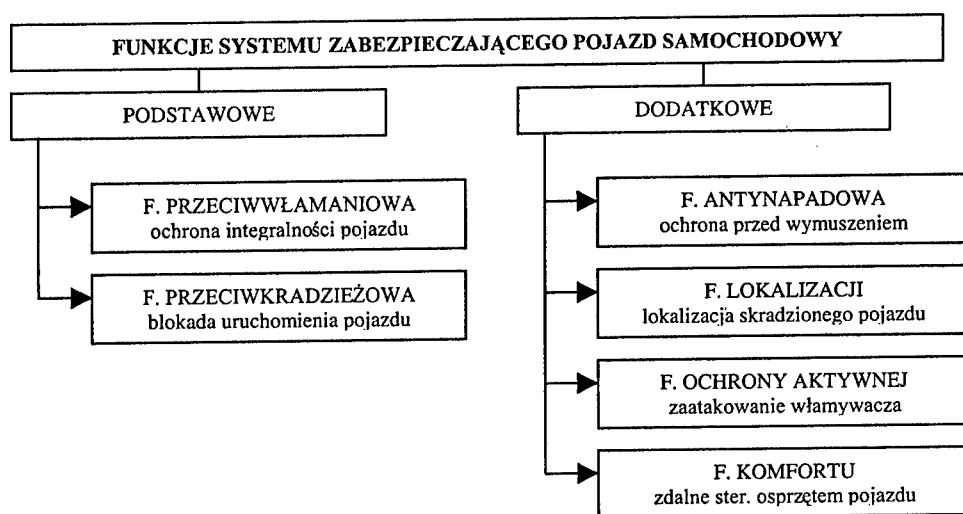
Streszczenie

W artykule przedstawiono budowę i ograniczenia konwencjonalnych samochodowych systemów zabezpieczających (SSZ) przed włamaniem i kradzieżą. Systemom obecnym na rynku przeciwstawiono nową koncepcję rozproszonego systemu zabezpieczającego zbudowanego z modułów połączonych magistralą LIN, dedykowaną do zastosowań w systemach elektroniki samochodowej. Przedstawiono zalety nowego rozwiązania oraz prototyp systemu, zrealizowany w Katedrze Metrologii i Systemów Elektronicznych.

1. WSTĘP

Włamania i kradzieże pojazdów samochodowych są od początku lat osiemdziesiątych, zgodnie z zasadą „miecza i tarczy” motorem rozwoju elektronicznych systemów zabezpieczających pojazdy przed włamaniem i/lub kradzieżą. Rozwój samochodowych systemów zabezpieczających (SSZ) przebiegał odmiennie po obu stronach „żelaznej kurtyny”. W krajach Europy Zachodniej, gdzie największym zagrożeniem dla pojazdu była kradzież pojazdu w całości największy nacisk położono na funkcję przeciwkradzieżową systemu – zabezpieczenie pojazdu przed uruchomieniem przez osoby nieupoważnione. Na rynku krajowym, ze względu na niedobór części zamiennych (ogumienie, akumulatory) największe zagrożenie stanowiły włamania do pojazdów oraz kradzieże celem pozyskania części. W związku z tym od początku lat 80-tych powstawały elektroniczne zabezpieczenia przeciwwłamaniowe, wykrywające, a następnie sygnalizujące akustycznie i optycznie naruszenie integralności pojazdu. Skutki tych odmiennych dróg rozwoju widoczne są do dzisiaj: w krajach Unii Europejskiej najpopularniejszą formą ochrony elektronicznej pojazdu jest zamontowany fabrycznie Immobilizer (zabezpieczenie przeciwkradzieżowe), zgodny z wymaganiami UE [1]. Systemy przeciwwłamaniowe stanowią niewielki odsetek i głównie są to kosztowne urządzenia fabryczne (tzw. *Original Equipment OE*) montowane w pojazdach luksusowych. Na rynku krajowym przeważają urządzenia łączące funkcję przeciwwłamaniową i przeciwkradzieżową – popularnie nazywane alarmami – niemal w 100% pochodzące od niezależnych producentów (tzw. *Aftermarket AM*). Ciekawostką

jest fakt, że analogiczna do rynku krajowego popularność systemów przeciwwłamaniowych AM występuje w Wielkiej Brytanii – kraju zmagającym się z plagą kradzieży samochodów. Wymagania dotyczące SSZ wydawane przez brytyjskie zrzeszenie ubezpieczycieli, tzw. Thatcham [2] stanowią często punkt odniesienia przy ocenie SSZ. Funkcje realizowane przez współczesne SSZ przedstawia rys. 1.



Rys.1. Funkcje realizowane przez samochodowy system zabezpieczający

2. BUDOWA I OGRANICZENIA KONWENCJONALNYCH SYSTEMÓW ALARMOWYCH

Urządzenia dominujące na rynku krajowym służą głównie do zabezpieczenia pojazdu przed włamaniem oraz dodatkowo pełnią funkcję przeciwwłamaniową, blokując jeden obwód układu wtryskowo-zapłonowego pojazdu. Produkty pochodzące od różnych producentów, konstrukcyjnie są bardzo zbliżone i oferują podobny poziom zabezpieczenia.

2.1 Budowa konwencjonalnego SSZ

Typowy SSZ zbudowany jest z centrali systemu, elektronicznego sygnalizatora akustycznego (syreny), diody LED sygnalizującej stan systemu, nadajników radiowych zdalnego sterowania (tzw. „pilotów”) i czujnika ruchu wewnątrz pojazdu. Centrala alarmowa jest najważniejszym elementem systemu. Sygnały wejściowe centrali to: zasilanie z instalacji elektrycznej pojazdu, sygnał włączenia stacyjki pojazdu, sygnały z wyłączników krańcowych drzwi i pokryw, sygnały z zewnętrznych czujników (czujnik położenia, czujnik tłuczonej szyby, czujnik uderzeniowy, czujniki ruchu) i ew. przewody od nadajnika i odbiornika ultradźwiękowego czujnika ruchu wbudowanego w centralę.

Sygnały wyjściowe centrali to wyjścia sygnalizacji optycznej (kierunkowskazy), wyjście sygnalizacji akustycznej (syrena elektroniczna), wyjścia sterowania z pilota alarmu systemem centralnego zamykania drzwi i domykania elektrycznych podnośników szyb pojazdu, zaciski przełączników rozwierających obwody systemów wtryskowo-zapłonowych pojazdu, wyjścia sterujące urządzeniami powiadamiania właściciela o stanie alarmu,

wyjścia potwierdzające uzbrojenie systemu i załączające czujniki dodatkowe oraz wyjścia załączania dodatkowych urządzeń z nadajnika zdalnego sterowania systemu (tzw. kanał 2).

Syrena systemu alarmowego sygnalizuje stan naruszenia pojazdu z maksymalną, dopuszczaną przez prawo głośnością [1][2][3]. W przypadku odcięcia od instalacji alarmowej (lub utraty zasilania) syrena włącza sygnalizację akustyczną. Syreny posiadają stacyjkę, umożliwiającą odłączenie zasilania systemu (demontaż akumulatora) bez wyzwiania sygnalizacji akustycznej. Niestety, obecność stacyjki nie pozwala prawidłowo ukryć syreny przed dostępem osób trzecich. Oprócz usunięcia stacyjki, pożądane byłoby potwierdzanie włączania i wyłączania systemu z ograniczoną głośnością oraz doładowywanie wewnętrznych akumulatorów syreny wyłącznie w czasie pracy silnika, ze względu na tendencje do zmniejszania dopuszczalnego poboru prądu przez SSZ [2].

2.2. Ograniczenia wynikające z budowy konwencjonalnego SSZ

Centralka systemu powinna być zamontowana w pojeździe w sposób utrudniający dostęp do niej podczas próby kradzieży. Wiązki przewodów instalacji alarmowej powinny być możliwie trudne do znalezienia w pojeździe, tak by utrudnić znalezienie centralki na podstawie układu wiązek. Zalecenie powyższe jest trudne do spełnienia ze względu na dużą liczbę przewodów doprowadzanych do centralki, tworzących grube i sztywne wiązki. Trudno jest znaleźć miejsce montażu centralki, w którym możliwe byłoby jednocześnie ukrycie wiązek przewodów. Niekiedy centralki wymagają dostępu serwisowego celem regulacji wbudowanego czujnika ruchu, co wymusza „płytki” montaż w pojeździe. Wbudowany w centralkę alarmu odbiornik radiowy nie pozwala montować urządzenia za metalowymi elementami deski rozdzielczej. Bezpośrednie połączenie diody LED i nadajników ultradźwiękowego detektora ruchu z centralką ułatwia jej znalezienie poprzez śledzenie przebiegu przewodów od w/w elementów oraz umożliwia atak generatorami wysokiego napięcia. Przewody silnoprądowe (do 20A) przekaźników unieruchamiających pojazd powinny mieć jak najmniejszą rezystancję (długość), co nie pozwala montować centralki daleko od deski rozdzielczej. Zwarcie tych przewodów w dowolnym miejscu wiązki lub przy centralce pozwala uruchomić pojazd (odłącza zabezpieczenie przeciwkradzieżowe systemu). Dodatkowo występują trudności montażowe wynikające z przeznaczenia urządzeń klasy AM do wielu typów samochodów. Instalację zabezpieczającą łączy się z instalacją oryginalną pojazdu w wielu miejscach. Umieszczenie punktów łączenia instalacji zależy od modelu, rocznika i kompletacji wyposażenia danego egzemplarza pojazdu. Niezbędne jest przygotowywanie przez instalatora dedykowanych wiązek, co wydłuża i utrudnia montaż urządzenia. Rozszerzenie funkcjonalności syreny o cechy wymienione w pkt. 2.1 powoduje zwiększanie ilości przewodów we wiązce syreny, co znacząco utrudnia prowadzenie wiązek przez przepusty grodzi przedniej pojazdu.

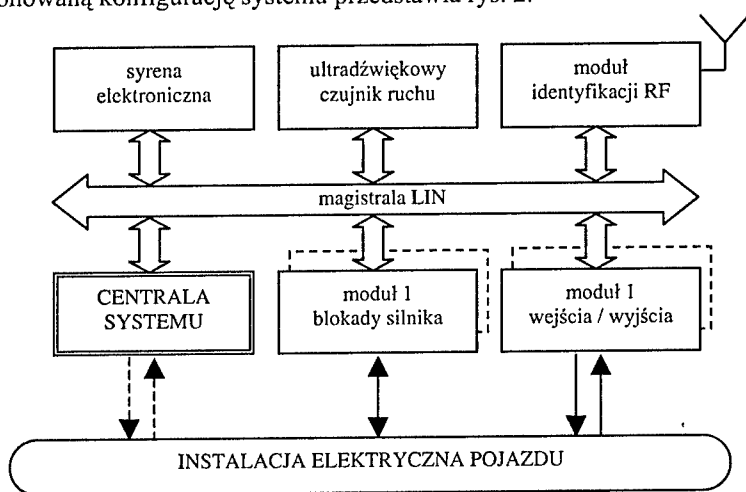
Jak wykazano powyżej, systemy konwencjonalne posiadają szereg wad. Przy pomocy tego typu systemów coraz trudniej będzie spełnić zarówno obecne, jak też projektowane wymagania ubezpieczycieli Europejskich, np. związane z obowiązkiem stosowania kodowanych sygnałów pomiędzy komponentami systemu i odpornością na sabotaż [3].

3. KONCEPCJA ROZPROSZONEGO SYSTEMU ZABEZPIECZAJĄCEGO

Stworzenie systemu zabezpieczającego nie posiadającego wad opisanych powyżej jest możliwe przy zastosowaniu rozwiązań, które sprawdziły się w systemach komputerowych, systemach pomiarowych oraz systemach automatyki przemysłowej. Zastosowanie moduło-

wej konstrukcji, dedykowanych magistral cyfrowych, interfejsów i protokołów do sterowania i wymiany danych w systemie, rozdzielenie modułu decyzyjnego od modułów linii wejścia/wyjścia, umożliwienie konfigurowalności i skalowalności systemu pozwoliłoby realizować systemy bardziej odporne na próby sabotażu, łatwiejsze do instalacji w pojeździe oraz bardziej uniwersalne od rozwiązań konwencjonalnych.

Proponowana koncepcja systemu polega na rozdzieleniu zadań pomiędzy moduły połączone dedykowaną magistralą cyfrową. Centralka systemu, zrealizowanego według przedstawionej koncepcji oprócz algorytmów zbliżonych do konwencjonalnych systemów przeciwwłamaniowych, musi posiadać mechanizmy związane z konfiguracją i zapewnieniem bezpieczeństwa transmisji danych pomiędzy modułami systemu. Niezbędna jest ciągła wzajemna autoryzacja modułów wykonawczych i centralki systemu oraz zapewnienie możliwości swobodnego przypisywania funkcji do linii modułów wejścia/wyjścia. Zaproponowaną konfigurację systemu przedstawia rys. 2.



Rys.2. Schemat blokowy rozproszonego systemu zabezpieczającego

3.1. Moduły wejścia/wyjścia

Moduły wejścia/wyjścia przeznaczone są do montażu w miejscach, gdzie występuje koncentracja sygnałów potrzebnych do pracy systemu alarmowego, np. okolice skrzynki bezpieczników, fabrycznej centralki systemu komfortu lub fabrycznych koncentratorów instalacji multipleksowanej. Połączenia modułów z fabryczną instalacją pojazdu są wykonywane krótkimi odcinkami przewodów. Dwa lub trzy takie moduły mogą monitorować i generować wszystkie sygnały instalacji alarmowej. Dodatkowo, dla uproszczenia struktury, funkcję modułu wejścia/wyjścia mogłaby pełnić także centralka systemu. Połączenie modułów z centralką są wykonywane przy pomocy kilkużyłowego, cienkiego przewodu i byłyby łatwe do wykonania i ukrycia w pojeździe.

3.2. Moduły blokad

Szczególnym przypadkiem modułu wejścia/wyjścia jest moduł blokady – przekaźnik rozłączający obwód (obwody) układu wtryskowo-zapłonowego pojazdu. Moduły blokady, o możliwie małych rozmiarach można umieszczać blisko newralgicznych obwodów

pojazdu: elektrycznej pompy paliwa, rozrusznika, sterownika silnika. Zastosowanie krótkich połączeń obwodów blokowanych znacząco zwiększa odporność systemu na sabotaż – w przeciwieństwie do konwencjonalnych rozwiązań, znalezienie centrali lub wiązki systemu nie pozwala „mostkować” obwodów blokowanych; ponadto system może posiadać kilka blokad, umieszczonych zarówno wewnątrz przedziału pasażerskiego jak też pod maską silnika pojazdu.

3.3. Moduł ultradźwiękowego czujnika ruchu

Moduł ultradźwiękowego czujnika ruchu jest przeznaczony do montażu blisko nadajnika i odbiornika ultradźwięków. Połączenie magistralą z systemem alarmowym, pozwalała regulować zdalnie czułość bez konieczności fizycznego dostępu do czujnika. Moduł ultradźwiękowego czujnika ruchu posiada linie wejściowe ze względu na prawdopodobieństwo montażu blisko lampki oświetlenia wnętrza i możliwość pozyskania sygnału otwarcia drzwi pojazdu.

3.4. Moduł identyfikacji użytkownika

Podstawową metodą identyfikacji użytkownika w systemach przeciwwłamaniowych jest transmisja zmiennej sekwencji kodowej przez nadajniki zdalnego sterowania tzw. „piloty”. Rozwiązania najpopularniejsze wykorzystują transmisję radiową 433.92 MHz, choć w systemach starszych lub o podwyższonym poziomie bezpieczeństwa spotkać można transmisję w podczerwieni. W przypadku podczerwieni, niezbędne jest umiejscowienie odbiornika w widocznym miejscu. W przypadku transmisji radiowej, optymalne jest umieszczenie odbiornika z dala od elementów blaszanych, powyżej linii okien pojazdu. Rozdzielenie układu identyfikacji od centrali i modułów we/wy pozwala zarówno uzyskać lepszy zasięg „pilotów” jak też lepiej ukryć pozostałe komponenty (centralę) systemu. Ponadto, możliwe jest stosowanie różnych metod identyfikacji (podczerwień, radio, transponder), w zależności od pożądanej konfiguracji systemu.

3.5. Sygnalizator akustyczny

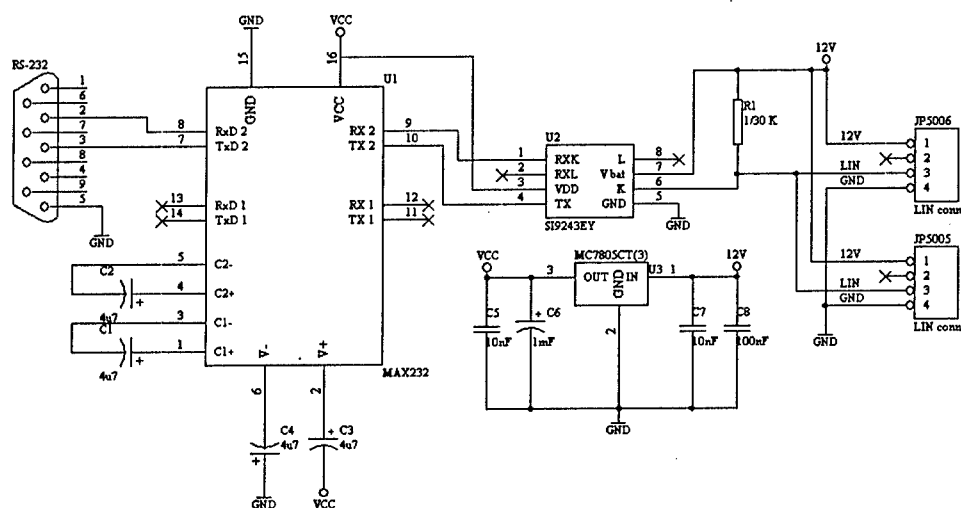
Zastosowanie magistrali cyfrowej łączącej syrenę z centralą pozwoli wyeliminować stacyjkę syrenki oraz wprowadzić funkcje opisane w pkt. 2 przy jednoczesnej redukcji ilości przewodów łączących syrenę z centralą. Dodatkowo, urządzenie powinno pozwalać indywidualizować system alarmowy – analogicznie jak w przypadku telefonów komórkowych – oferować wybór dźwięków sygnalizacji alarmowania i zmiany stanu systemu oraz możliwość stosowania komunikatów głosowych. Sygnalizator akustyczny powinien posiadać także linię wejściową przeznaczoną do dołączenia wyłącznika krańcowego pokryw silnika, znajdującego się zazwyczaj blisko syreny.

4. REALIZACJA LABORATORYJNA ROZPROSZONEGO SYSTEMU ZABEZPIECZAJĄCEGO

Wykonano model umożliwiający badania i prezentację w laboratorium dydaktycznym idei rozproszonego systemu zabezpieczającego. Opis modelu znajduje się w pracy [4].

4.1. Magistrala LIN

Zdecydowano się na zastosowanie magistrali LIN – Local Interconnect Network – do wymiany danych pomiędzy modułami systemu. Ze względu na proste okablowanie (jedenprzewodowa warstwa fizyczna interfejsu) i niski koszt realizacji pojedynczego węzła, magistrala LIN stanowi obecnie na rynku motoryzacyjnym tańszą alternatywę dla magistrali CAN w systemach, w których nie jest wymagana duża prędkość transmisji danych (2400-19200 bit/s). Magistrala LIN jest hierarchiczna: posiada jeden węzeł nadrzędny typu Master (w centrali systemu) oraz wiele węzłów typu Slave. Ramkę transmisji LIN można realizować przy pomocy dedykowanych sterowników LIN lub z wykorzystaniem prostych interfejsów UART wbudowanych we większość współczesnych mikrokontrolerów. Magistrala zapewnia synchronizację Mastera i Slave'ów, dzięki czemu mikrokontrolery urządzeń Slave mogą pracować z rezonatorami RC. Zalecane jest, by magistrala nie była dłuższa niż 40m, co odpowiada dziesięciokrotnej długości typowego pojazdu. Zaleca się także, by liczba węzłów magistrali nie przekraczała 16. Istnieją dedykowane układy scalone do realizacji warstwy fizycznej interfejsu, zawierające stopień wyjściowy z otwartym drenem, zabezpieczenia przeciwprądowe i przeciwprzepięciowe. Ze względu na dostępność elementów, w pracy wykorzystano układy warstwy fizycznej pochodzące od samochodowej magistrali diagnostycznej ISO9141 o parametrach elektrycznych zbliżonych do magistrali LIN. Schemat konwertera poziomów z RS232 na LIN przedstawia rys. 3.



Rys. 3. Konwerter RS232-LIN

4.2. Model rozproszonego systemu zabezpieczającego

Zrealizowane zostały następujące moduły rozproszonego SSZ: sygnalizator akustyczny z funkcją komunikatów głosowych, ultradźwiękowy detektor ruchu, moduł wejścia/wyjścia, i moduł identyfikacji użytkownika. Moduły połączono równolegle za pomocą trzech przewodów: zasilania, masy i linii danych. Zrealizowano część sprzętową i opro-

gramowanie w/w modułów. Pracą każdego z modułów steruje tani, 8-bitowy mikrokontroler Atmel z rodziny AVR, pełniący funkcję sterownika interfejsu LIN. Oprogramowanie mikrokontrolerów zawartych w modułach systemu zostało napisane w języku C, w środowisku AVR Studio, z wykorzystaniem kompilatora AVR GCC.

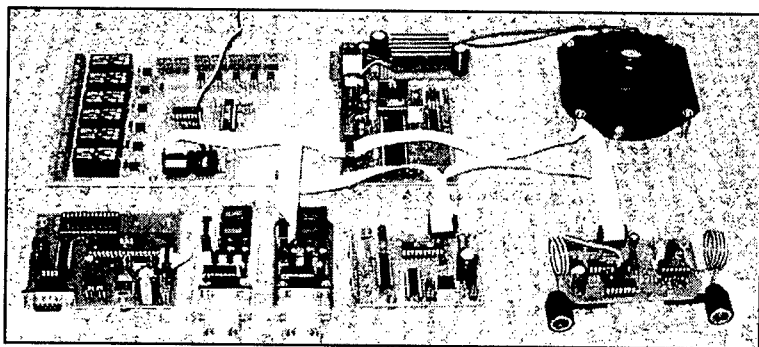
Najciekawszym modułem wykonanego modelu jest sygnalizator akustyczny z funkcją odtwarzania komunikatów głosowych. Dźwięk wytwarzany jest za pomocą generatora PWM wbudowanego w mikrokontroler. Komunikaty zapisane są w postaci cyfrowej w pamięci nieulotnej typu DataFlash. Możliwe jest przygotowanie i edycja zarówno dźwięków, jak też komunikatów głosowych na komputerze PC, a następnie zapisanie rekordów sygnałów zdalnie przez magistralę LIN do systemu plików w pamięci sygnalizatora akustycznego.

Zrealizowano jeden moduł wejścia/wyjścia zawierający linie wejściowe, wyjściowe i przekaźniki blokad. Wyłączniki krańcowe drzwi pojazdu oraz diody LED sygnalizujące obwody świateł i kierunkowskazów zamontowano w miniaturowym modelu pojazdu.

Funkcje centralki sterującej systemem pełni komputer PC z dołączonym konwerterem RS232-LIN. Oprogramowanie sterujące systemem zostało napisane w środowisku LabWindows/CVI. Opracowano wstępną wersję oprogramowania centralki systemu, realizującą algorytm pracy systemu zabezpieczającego oraz umożliwiającą konfigurację systemu i przypisanie funkcji do linii modułu wejścia/wyjścia.

4.3. Badania przeprowadzone na modelu systemu

Ze względu na charakter modelu przeznaczonego do testowania koncepcji systemu, przeprowadzono jedynie badania funkcjonalne. Sprawdzone, czy system realizuje zaprojektowany algorytm nadzoru pojazdu samochodowego. System można uzbroić i rozbroić przy pomocy nadajnika zdalnego sterowania. Po uzbrojeniu otwarcie drzwi modelu lub naruszenie strefy chronionej przez czujnik ruchu powoduje sygnalizację stanu alarmu. Dodatkowo, zaimplementowano dwie funkcje dodatkowe – zdalne załączanie świateł pojazdu oraz odtwarzanie komunikatu ostrzegawczego po naciśnięciu odpowiednich przycisków nadajnika zdalnego sterowania.



Rys. 4. Model rozproszonego systemu zabezpieczającego

Zmierzono czas reakcji systemu na zdarzenia. Czas reakcji na zmianę stanu dozorowanych linii wynikał z przyjętego algorytmu weryfikacji stabilności stanu linii i wynosił poniżej 100 ms. Zbadano czas od naciśnięcia pilota przez użytkownika do potwierdzenia odebrania komunikatu za pomocą sygnalizatorów akustycznych i optycznych. Czas ten

zależy od czasu transmisji sekwencji kodowej pilota, czasu dekodowania sygnału pilota przez moduł identyfikacji, czasu potrzebnego na poinformowanie modułu centralnego o naciśnięciu pilota przez moduł identyfikacji, czasu pracy algorytmu sterującego systemem oraz czasu potrzebnego na wysłanie przez moduł centralny rozkazów do modułu wykonawczego. Doświadczalnie zweryfikowano, że nawet najmniejsza prędkość transmisji magistrali LIN (2400bps) pozwala na zachowanie czasu reakcji systemu na pilota poniżej 500ms, co gwarantuje komfortowe dla użytkownika „sprężenie zwrotne” systemu na obsługę pilota.

5. ZAKOŃCZENIE

Praktyczna realizacja systemu pozwoliła sprawdzić w laboratorium działanie rozproszonego systemu zabezpieczającego. Zbudowany model potwierdził możliwość stosowania tanich i prostych mikrokontrolerów jako węzłów magistrali LIN, a także możliwość realizacji sygnalizatora akustycznego z cyfrowym zapisem komunikatów. Urządzenie komercyjne, zrealizowane w/g tego pomysłu byłoby łatwe w montażu, bardziej odporne na sabotaż od rozwiązań konwencjonalnych oraz diagnozowalne przy pomocy komputera. Niestety, wymagałoby od instalatora pewnych kwalifikacji oraz dostępu do komputera PC podczas montażu i konfiguracji systemu.

BIBLIOGRAFIA

- [1] Dyrektywa 95/56/EC: *Commission directive 95/56/EC of 8 November 1995 adapting to technical progress Council Directive 74/61/EEC relating to devices to prevent unauthorised use of vehicles*, CONSLEG 1995L0056.
- [2] Thatcham, the motor insurance repair research centre, *The British Insurance Industry's Criteria for Vehicle Security*. Issue 2: June 1996
- [3] Stichting Certificering Motorrijtuig beveiliging, *Homologation directive AA03 Electronic security for passenger cars*, SCM, Holandia 1996
- [4] Olejniczak M., *Rozproszony system zabezpieczający pojazd samochodowy*, Praca dyplomowa WETI 2003

CONCEPT OF MODULAR VEHICLE SECURITY SYSTEM BASED ON DEDICATED AUTOMOTIVE BUS

Summary

The paper presents fundamentals and limitations of conventional (single unit) Vehicle Security Systems. To overcome these limitations, the new concept of modular Vehicle Security System is presented, based on LIN bus, dedicated for low transmission rate automotive industry solutions. Benefits of new concept are presented, together with realization of laboratory model of such system.

Jerzy Pluciński, Paweł Wierzbą

Katedra Optoelektroniki, Politechnika Gdańska

OPTOELEKTRONICZNE METODY OCHRONY INFRASTRUKTURY TELEINFORMATYCZNEJ

Streszczenie

Kluczowe znaczenie infrastruktury teleinformatycznej w funkcjonowaniu współczesnego państwa wymaga zapewnienia bezpieczeństwa jej funkcjonowania. W referacie omówiono metody kontroli dostępu do obiektów infrastruktury teleinformatycznej wykorzystujące optyczne techniki wykrywania i zobrażenia oraz interferometryczne rozłożone sensory światłowodowe i sieci sensorowe. Przedstawiono zastosowanie technik biometrycznych w identyfikacji i autoryzacji użytkowników. Zaprezentowano najnowsze metody optoelektronicznego monitorowania stanu łączy światłowodowych wykorzystujące reflektometrię światłowodową oraz unikatowe metody optycznej transmisji informacji umożliwiające wykrycie próby nieautoryzowanego odczytu informacji.

1. WSTĘP

Współczesne sieci teleinformatyczne są istotnym elementem infrastruktury umożliwiającej poprawne funkcjonowanie podmiotów gospodarczych, instytucji i agend rządowych. Stąd też zapewnienie poprawnej pracy tych sieci jest zadaniem najwyższego priorytetu.

Podstawowymi zagrożeniami dla poprawnego funkcjonowania sieci teleinformatycznych są: uszkodzenia elementów infrastruktury teleinformatycznej spowodowane zjawiskami naturalnymi (np. klęski żywiołowe) lub działaniem celowym (akty wandalizmu lub sabotażu), nieautoryzowany dostęp do danych w czasie ich wprowadzania, przesyłania lub przechowywania.

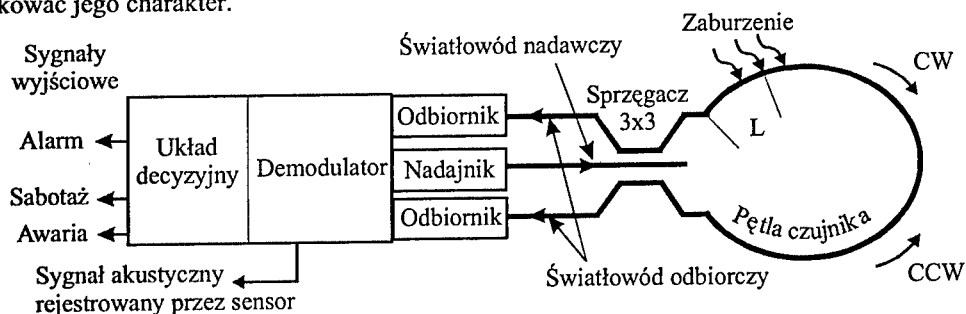
Przeciwdziałanie tym zagrożeniom wymaga: (a) kontroli dostępu do elementów infrastruktury, (b) monitorowania stanu łączy światłowodowych, (c) identyfikacji i autoryzacji osób mających dostęp do infrastruktury. Przesłanką przemawiającą za szybkim podjęciem wymienionych działań jest przewidywany, znaczny wzrost liczby prób uzyskania dostępu do danych przesyłanych w infrastrukturze teleinformatycznej, wynikający z jej gwałtownego rozwoju, oraz z obserwowanej poprawy stanu zabezpieczeń systemów komputerowych (uzyskanie dostępu do danych w czasie ich przesyłania staje się łatwiejsze niż włamanie do systemów komputerowych, w których są one przechowywane).

2. KONTROLA DOSTĘPU DO ELEMENTÓW INFRASTRUKTURY TELEINFORMATYCZNEJ

Kontrola dostępu do elementów infrastruktury teleinformatycznej ma na celu przeciwdziałanie aktom wandalizmu lub sabotażu oraz uniemożliwienie nieautoryzowanego dostępu do przesyłanych danych. Dozorowane urządzenia, takie jak np. optyczne wzmacniacze sygnału lub stacje bazowe telefonii komórkowej, są często zlokalizowane w miejscach trudno dostępnych lub na terenach niezamieszkałych. Powoduje to konieczność nadzoru nie tylko miejsca instalacji elementów infrastruktury, ale także jego otoczenia – strefy ochronnej obiektu, tak by możliwa była skuteczna reakcja na próbę uzyskania nieautoryzowanego dostępu.

Stosowane obecnie systemy zabezpieczeń wykorzystują m.in. pojemnościowe sensory zbliżeniowe, sensory ruchu pracujące w podczerwieni, sensory drgań i wibracji, elektryczne i światłowodowe sensory nacisku oraz kamery pracujące w świetle widzialnym. Problemem występującym w tych systemach jest niewystarczająca selektywność stosowanych w nich sensorów (alarmy wywoływane przez ptaki lub zwierzęta) oraz ograniczony obszar nadzorowany przez jeden sensor. Zastosowanie kamer poprawia selektywność systemów, ale wymaga ciągłego nadzoru przez operatora. Ponadto ich skuteczność zależy od warunków atmosferycznych (mgła, opady) i pory dnia.

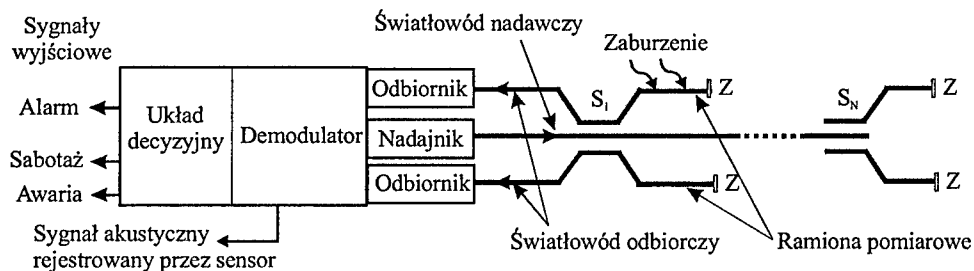
Celem prac nad sensorami światłowodowymi i nowymi konstrukcjami detektorów podczerwieni jest opracowanie rozwiązań wolnych od powyższych wad. Jednym z obiecujących rozwiązań jest sensor światłowodowy wykorzystujący interferometr pętlowy. Sensor ten, przedstawiony na Rys. 1, pozwala na monitorowanie strefy ochronnej długości do 10 km, umożliwia lokalizację miejsca zaburzenia i pozwala określić jego charakter (pojazd, człowiek, zwierzę) [1]. W sensorze tym światło z nadajnika (lasera) jest rozdzielane w sprzęgaczu 3x3 na dwie wiązki poruszające się w pętli światłowodowej, umieszczonej wzdłuż brzozy strefy chronionej, w kierunku ruchu wskazówek zegara (CW) i w kierunku przeciwnym do kierunku ruchu wskazówek zegara (CCW). Wiązki te przechodząc przez obszar, w którym na światłowód działa zaburzenie, doznają przesunięcia fazy odpowiednio o $\Delta\phi(t)$ i $\Delta\phi(t+\tau)$ w odstępie czasu τ , zależnym do długości pętli i od odległości L tego obszaru od początku pętli. Następnie są one dodawane koherentnie w sprzęgaczu i doprowadzane światłowodami odbiorczymi do dwóch kanałów odbiornika. W wyniku zastosowania sprzęgacza 3x3 można określić znak różnicy faz interferujących wiązek [2], co pozwala zlokalizować miejsce działania zaburzenia, a w połączeniu z odpowiednimi metodami obróbki sygnałów także wyznaczyć jego widmo częstotliwościowe, a tym samym zidentyfikować jego charakter.



Rys. 1. Rozłożony sensor światłowodowy wykorzystujący interferometr pętlowy

Kolejnym interesującym rozwiązaniem jest przedstawiony na rys. 2. pseudorozłożony sensor interferometryczny [3] składający się z N światłowodowych interferometrów Michelsona o długości ramion L_M połączonych szeregowo odcinkami światłowodu o długości $L_0 > L_M$. W sensorze tym nadajnik (laser półprzewodnikowy) wytwarza impuls, którego długość L w światłowodzie jest mniejsza od L_0 . Przy przejściu przez kolejne sprzęgacze S_i część mocy impulsu jest odsprężana do każdej pary ramion pomiarowych kolejnych interferometrów. Tak powstałe nowe impulsy propagują dalej w ramionach pomiarowych, odbijają się od zwierciadeł Z i powracają do sprzęgacza. Na skutek działania zaburzenia, pomiędzy impulsami powstaje różnica faz $\Delta\phi(t)$. W sprzęgaczu oba impulsy interferują, a natężenie otrzymanego promieniowania jest funkcją tej różnicy.

W wyniku złożenia sygnałów pochodzących od kolejnych interferometrów otrzymywany jest sygnał będący ciągiem N impulsów o amplitudach zmiennych w czasie, umożliwiające określenie, na który z interferometrów działa zaburzenie. Wyznaczenie widma częstotliwościowego zmian fazy mierzonej przez poszczególne interferometry pozwala na identyfikację źródła tego zaburzenia.

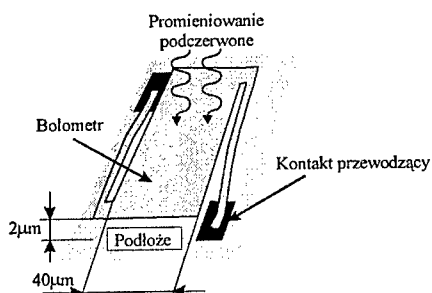


Rys. 2. Pseudorozłożony sensor światłowodowy wykorzystujący interferometry Michelsona

Celem prac nad detektorami stosowanymi w systemach zabezpieczeń jest opracowanie tanich i prostych rozwiązań o skuteczności niezależnej od warunków atmosferycznych i pory dnia. Prace te koncentrują się na detektorach pracujących w zakresie średniej podczerwieni – w zakresie długości fal od $3\ \mu\text{m}$ do $11\ \mu\text{m}$. Dotychczas większość detektorów stosowanych w tym zakresie widmowym wymagała chłodzenia termoelektrycznego do temperatury -35°C lub chłodzenia ciekłym azotem do -196°C , co uniemożliwiało ich zastosowanie w omawianych systemach.

Wraz z rozwojem technologii mikromechaniki (ang. *Micro-Opto-Electro-Mechanical Systems – MOEMS*) opracowuje się matryce niechłodzonych detektorów podczerwieni. Detektorami stosowanymi w matrycach są zwykle bolometry (przykładową realizację piksela takiej matrycy, produkcji firmy Honeywell przedstawiono na rys. 3) Bolometry te absorbują promieniowanie wyemitowane przez obserwowany obiekt lub odbite od niego, co powodując wzrost ich temperatury. Ponieważ bolometry wykonane są z materiału o dużym temperaturowym współczynniku rezystancji (TWR), zmiany ich temperatury powoduje zmiany ich rezystancji, które następnie są przetwarzane na sygnał napięciowy. Sygnał ten dostarczany jest na wyjście matrycy przy pomocy układów podobnych do stosowanych w detektorach CCD czy CMOS.

Ponieważ detektory bolometryczne wykorzystują najczęściej promieniowanie wytworzone przez obserwowany obiekt, mogą one pracować bez konieczności stosowania oświetlenia terenu. Ponadto promieniowanie podczerwone jest znacznie słabiej tłumione przez opady, mgłę czy zadymienie, co w znacznym stopniu uniezależnia skuteczność tej metody od warunków atmosferycznych.



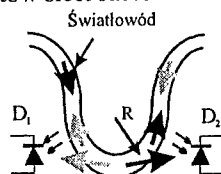
Rys. 3. Bolometr wykonany w technologii mikromechaniki [4]

Omawiane matryce dostępne są w rozdzielczościach od 160x240 pixeli do 640x480 pixeli i charakteryzują się szybkością pracy od 10 do 30 ramek/sekundę, co jest wartością wystarczającą w systemach zabezpieczeń. Obecnie ceny kamer wykorzystujących niechłodzone detektory bolometryczne kształtują się na poziomie kilku tysięcy USD i szybko spadają, co powoduje znaczący wzrost zainteresowania

3. MONITOROWANIE STANU ŁĄCZY ŚWIATŁOWODOWYCH

Monitorowanie stanu łączy światłowodowych ma na celu ich zabezpieczenie przed próbami uzyskania nieautoryzowanego dostępu do danych w czasie ich przesyłania oraz lokalizację miejsca uszkodzeń kabli światłowodowych. Jak dotąd zagadnieniu monitorowania łączy światłowodowych nie poświęcano zbyt wiele uwagi, pomimo że wypadki prowadzenia podsłuchu tego typu łączy przez wywiady niektórych państw są ogólnie znane. Powodem tego był z jednej strony bardzo ograniczony dostęp do urządzeń umożliwiających uzyskanie dostępu do transmitowanych danych, z drugiej zaś wysoki poziom bezpieczeństwa osiągany dzięki zastosowaniu algorytmów szyfrujących z kluczem publicznym.

Obecnie, wraz z pojawieniem się szerokiej klasy aplikacji, takich jak *mBanking*, pracujących na platformach mobilnych (telefony komórkowe, palmtopy) o ograniczonej mocy obliczeniowej i mniej bezpiecznym szyfrowaniu, można obawiać się lawinowego wzrostu liczby prób uzyskania dostępu do danych transmitowanych łącami światłowodowymi do stacji bazowych lub innych elementów sieci radiokomunikacji ruchomej.



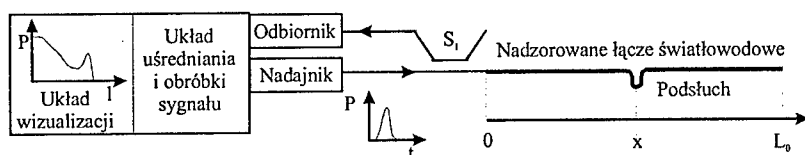
Rys. 4. Detekcja sygnału optycznego wypromieniowanego na zgięciu światłowodu

Dostęp taki można w najprostszym wypadku uzyskać wykonując zgięcie światłowodu o promieniu R wynoszącym kilka mm. Na zgięciu tym dochodzi do wypromieniowania ze światłowodu części prowadzonej w nim mocy promieniowania optycznego w sposób pokazany na rys. 4. Promieniowanie to padając na detektor jest zamieniane na sygnał elektryczny, który po wzmocnieniu i obróbce jest rejestrowany przez komputer. Sposób ten jest stosowany w urządzeniach do serwisowania łączy światłowodowych [5] i w spawarkach światłowodowych. Uzyskanie dostępu do włókna światłowodowego umieszczonego w kablu nie jest trudne, a ryzyko jego wykrycia jest niewielkie, ponieważ kable światłowodowe

nie zawierają w swojej konstrukcji żadnych elementów pozwalających na wykrycie naruszenia ich integralności.

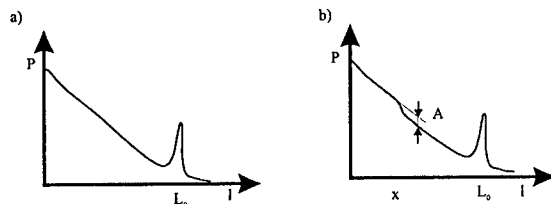
Należy jednak pamiętać, że w wyniku wypromieniowania części mocy ze światłowodu następuje zmniejszenie mocy sygnału docierającego do odbiornika, a tym samym pogorszenie stosunku sygnał/szum i wzrost stopy błędów (ang. *Bit Error Rate* – *BER*). Pomiar stopy błędów nie może jednak być wykorzystany do zabezpieczenia łącza przed próbami nieautoryzowanego dostępu. Systemy transmisji danych pracują bowiem z pewnym zapasem mocy, by zminimalizować wpływ powolnych zmian parametrów poszczególnych elementów łącza. Dlatego też zmiana stopy błędów wywołana przez wprowadzenie w tor światłowodowy zgięcia może być trudna do wykrycia.

Urządzeniem umożliwiającym wykrycie zgięcia wprowadzonego w tor światłowodowy jest światłowodowy reflektometr pracujący w dziedzinie czasu (ang. *Optical Time-Domain Reflectometer* – *OTDR*), przedstawiony na rys. 5. Do nadzorowanego łącza światłowodowego wprowadzany jest krótki (kilku nanosekundowy) impuls promieniowania podczerwonego. Podczas jego propagacji w światłowodzie część jego energii jest rozpraszana wstecznie (rozpraszanie Rayleigha). Kilka procent mocy impulsu odbija się od końca łącza (odbicie Fresnela). Sygnał odbity jest doprowadzany do detektora przez sprzęgacz S_1 i po niezbędnej obróbce sygnał ten jest przedstawiany w postaci zależności mocy P promieniowania odbitego wyrażonej w decybelach od odległości l mierzonej od początku łącza, jak na rys. 5.



Rys. 5. Reflektometryczny nadzór łącza światłowodowego

Typowy sygnał uzyskiwany z reflektometru przedstawiono na Rys. 6a. W sygnale tym można zaobserwować jedynie jednostajny spadek poziomu sygnału odbitego wywołany tłumieniem światłowodu oraz odbicie Fresnela na końcu łącza ($l=L_0$). Wprowadzenie zgięcia w tor światłowodowy w odległości x od jego początku powoduje, że rejestrowany sygnał przybiera postać przedstawioną na Rys. 6b. Wartość tłumienia A , które może zostać zmierzone, jest mniejsza od 0,1 dB, co zapewnia prawie 100% skuteczność wykrywania podśluchu.



Rys. 6. Sygnał rejestrowany przez reflektometr przy braku podśluchu (a) i przy jego obecności (b)

Uzyskanie dostępu do większości łączy infrastruktury teleinformatycznej i rejestracja przesyłanego sygnału nie pozwala na odczytanie przesyłanych danych, ze względu na stosowanie szyfrowania. Najczęściej stosowane algorytmy wykorzystują metodę szyfrowania opartą o dwa wzajemnie uzupełniające się klucze: publiczny i prywatny. Proces szyfrowania opiera się na potęgowaniu przesyłanej informacji modulo n przez klucz publiczny, zaś

proces deszyfrowania – na potęgowaniu informacji zaszyfrowanej modulo n przez klucz prywatny. Choć proces szyfrowania i deszyfrowania jest bardzo szybki, łamanie szyfru jest operacją bardzo pracochłonną (liczba niezbędnych kroków obliczeniowych rośnie wykładniczo z długością klucza). Przykładowo, stosując komputer wykonujący 10^{12} operacji na sekundę, złamanie szyfru z kluczem publicznym złożonym z 150 cyfr wymaga 150 tysięcy lat obliczeń. Wydawałoby się, że tego typu algorytm szyfrowania jest bezpieczny. Jednak Peter W. Shor zauważył, że tzw. komputery kwantowe wymagałyby znacznie mniej operacji do rozkładu dużych liczb na czynniki pierwsze (liczba operacji rośnie potęgowo z długością klucza) [6]. Złamanie szyfru z kluczem publicznym złożonym z 150 cyfr przy ich pomocy wymagałoby tylko $5 \cdot 10^{10}$ kroków, co przy prędkości 10^{12} operacji na sekundę zajęłoby znacznie mniej niż 1 sekundę.

Choć komputery kwantowe, nawet w najprostszej formie, nie są dostępne na rynku, ich pojawienie jest tylko kwestią czasu (ich budowę zajmuje się wiele przodujących laboratoriów fizycznych) [7]. W chwili ich pojawienia stosowane obecnie metody szyfrowania będą musiały ulec całkowitym zmianom.

Wykorzystanie właściwości kwantowych światła w połączeniu z telekomunikacją optyczną oferuje znacznie pewniejsze metody zabezpieczenia informacji [8]. Wykorzystują one efekt splątania zwany także efektem EPR (od Alberta Einsteina, Borysa Podolskiego i Nathana Rosena, którzy analizowali te zjawisko). Do przesyłania informacji wykorzystywane są pary fotonów splątanych. Z każdej pary jeden foton otrzymuje nadawca (foton A'), a drugi odbiorca (foton A'') informacji. By przesłać wiadomość najpierw modulowana jest wiązka światła, w wyniku czego powstają fotony X o stanach zgodnych z bitami wiadomości (np. bitom 0 może odpowiadać jeden stan polaryzacji, a 1 drugi, ortogonalny stan polaryzacji). Następnie dokonywany jest „pomiar” korelacji między fotonami A' i X – co jest odpowiednikiem szyfrowania informacji w tradycyjnych systemach teleinformatycznych. Wynik tego „pomiaru” transmitowany jest dalej do odbiorcy, który dokonuje zmianę stanu swoich kopii fotonów A'' zgodnie z otrzymanymi wynikami „pomiaru”. W wyniku tego powstaje strumień fotonów o stanie identycznym jak fotonów X. Odczyt stanów tych fotonów równoważny jest z odczytem przesyłanej wiadomości. Warto zauważyć, że nie ma możliwości powstania dwóch kopii fotonów X, zatem nie jest możliwy podsłuch przesyłanej wiadomości.

Dotychczas w warunkach laboratoryjnych dokonano transmisji światłowodowej stanów splątanych na odległość pojedynczych kilometrów. Jak dotąd nie stwierdzono istnienia ograniczeń uniemożliwiających stosowanie tej metody w transmisji na duże odległości. Dlatego też w niedalekiej przyszłości można oczekiwać wprowadzenia systemów – transmisji wykorzystujących stany splątane.

4. METODY AUTORYZACJI UŻYTKOWNIKÓW

Powszechnie stosowane metody autoryzacji wykorzystują karty dostępu, numery PIN i hasła, a zatem identyfikują numer, kartę czy hasło, a nie osobę. Do identyfikacji osobistej mogą być stosowane metody biometryczne, uzupełniając lub zastępując dotychczasowe rozwiązania. Metody te wykorzystują indywidualne zróżnicowanie cech fizycznych i behawioralnych, takich jak: linie papilarnie, kształt twarzy czy dłoni, charakterystyczne cechy tęczy oka, pismo ręczne, mowa, sposób uderzania w klawisze, czy układ żył nóg. Eliminując konieczność pamiętania hasła oraz posiadania kart dostępu, systemy biometryczne zapewniają wysoki poziom wygody przy jednoczesnym wysokim stopniu bezpieczeństwa.

Istotną rolę w metodach biometrycznych pełni optoelektronika. Gwałtowny rozwój metod obrazowania oraz spadek cen urządzeń optoelektronicznych umożliwia coraz szersze wykorzystanie metod biometrycznych do autoryzacji dostępu do laboratoriów badawczych, bankomatów, sieci komputerowych, systemów alarmowych, zamków drzwiowych, kart procesorowych itd. Wartość sprzedanych urządzeń do kontroli biometrycznej wyniosła w 2003 roku 719 mln USD. Przewiduje się, że osiągnie ona w 2004 roku wartość 1,201 mld USD, a do roku 2008 aż 4,639 mld USD [9-11].

Obecnie stosowane metody optyczne wykorzystują rozpoznawanie linii papilarnych, kształtu dłoni i obrazu tęczówki oka. Prowadzone są także badania nad wykorzystaniem obrazu naczyń krwionośnych siatkówki oka oraz nad wykorzystaniem cech anatomicznych twarzy: obrazu termicznego i układu naczyń krwionośnych.

Najbardziej znaną metodą biometryczną jest metoda oparta na rozpoznawaniu linii papilarnych, szeroko stosowana od początku XX wieku. Podstawą identyfikacji odcisków palców jest układ tzw. minutii, czyli punktów, w których linie papilarne rozdwajają się, łączą lub kończą ślepym zaułkiem. Obraz linii papilarnych rejestrowany jest przy pomocy sensora optycznego. Z obrazu tego wyznaczany jest układ punktów charakterystycznych i innych cech identyfikujących, który jest następnie porównywany z zapamiętanym wzorcem. Identyfikacja za pomocą linii papilarnych ma swoje wady: głębsze uszkodzenia powierzchni skóry mogą mieć wpływ na odczyt, a więc i na identyfikację danej osoby.

Od wady jest w znacznym stopniu wolna metoda identyfikacji przy pomocy geometrii dłoni. W metodzie tej wykonywane jest trójwymiarowe zdjęcie dłoni, na podstawie którego określana jest długość, szerokość, grubość czterech palców oraz wielkość obszarów pomiędzy kostkami. Łącznie wykonywanych jest ponad 90 pomiarów różnych cech charakterystycznych dłoni. Wynik tych pomiarów jest przechowywany w pamięci urządzenia w formie (maksimum) 9 bajtowego wzorca. Identyfikator ten jest unikatowy dla każdego człowieka.

Jednym z najbardziej unikatowych identyfikatorów jest tęczówka oka. Kształtuje się ona w ciągu dwóch pierwszych lat naszego życia i ulega zniszczeniu w maksymalnie 5 sekund po zgonie, co uniemożliwia jej wykorzystanie jako wyizolowanej tkanki. Zawiera ona średnio 266 punktów charakterystycznych (parokrotnie więcej niż punktów charakterystycznych odcisku palca), które definiują cechy identyfikujące tęczówkę. Cechy te pozostają niezmiennie aż do śmierci. Poza mechanicznym uszkodzeniem i przypadkiem nowotworu nie zanotowano odstępstw od tej reguły. Obecne systemy rozpoznawania tęczówki nie są wrażliwe na przypadkowe i celowe ruchy głowy, mrugnięcie czy przymknięcie powieki. Rozpoznają one bezbłędnie osoby nawet wtedy, kiedy noszą one okulary lub szkła kontaktowe. W systemie identyfikacji przy pomocy tęczówki oka na początku wykonywane jest zdjęcie twarzy, na którym określane jest położenie oczu. Następnie specjalna kamera wykonuje zdjęcie tęczówki o bardzo wysokiej rozdzielczości. Na jego podstawie wyznaczany jest kod zawierający skrócony opis punktów charakterystycznych. Kod ten jest następnie szyfrowany (z zastosowaniem szyfrów jednokierunkowych) i porównywany z zapisanym w bazie danych zaszyfrowanym kodem oryginału. Porównywanie zaszyfrowanych kodów zapewnia bardzo wysoki stopień bezpieczeństwa, ponieważ nawet uzyskanie dostępu do zapisanego w bazie danych kodu nie umożliwia odtworzenia obrazu tęczówki. Istniejące rozwiązania zapewniają poziom błędów rzędu 10^{-10} (w niektórych bardziej zaawansowanych systemach poziom błędów może osiągać nawet wartość 10^{-20}). Dlatego też systemy tego typu wykorzystywane są coraz częściej w szczególnie wymagających zastosowaniach [12].

5. ZAKOŃCZENIE

Zaprezentowane w referacie metody zabezpieczenia infrastruktury teleinformatycznej umożliwiają poprawę bezpieczeństwa gromadzonych i przesyłanych w niej danych. Przesłanką przemawiającą za koniecznością stosowania omówionych metod jest przewidywany, znaczny wzrost liczby prób uzyskania dostępu do danych przesyłanych w infrastrukturze teleinformatycznej, wynikający z obserwowanej poprawy stanu zabezpieczeń systemów komputerowych. Zastosowanie prezentowanych metod może wyeliminować lub znacznie zmniejszyć prawdopodobieństwo uzyskania dostępu do przesyłanych danych.

BIBLIOGRAFIA

- [1] Jaroszewicz L.R.: *Rola polaryzacji i spójności w interferometrii światłowodowej*, WAT Warszawa, 1995, s.161.
- [2] Chiu B., Hastings M.C.: *Demodulation of output signals for passive homodyne optical fibre interferometry based on 3x3 coupler*. Proc. SPIE, vol. 2292, 1994, pp. 371-381.
- [3] Szustakowski M., Chojnacki M.: *Wielostrefowy czujnik światłowodowy z rozłożonym polem detekcji w monitorowaniu stref ochronnych*. w: VII Konferencja Naukowa „Czujniki Optoelektroniczne i Elektryczne” Rzeszów 5-8 czerwca 2002, t. 1, s. 91-96.
- [4] Wood R.A.: *Monolithic silicon microbolometer arrays*. w: Semiconductors and semimetals. Vol. 47. Academic Press 1997, pp. 45-121.
- [5] Fiber optic talk set type FST-1 with a FTS-20C clip-on coupler, prod. Alcoa Inc, http://www.alcoa.com/afl_tele/en/product.asp?cat_id=130&prod_id=202.
- [6] Vandersypen L.M.K., Steffen M., Breyta G., Yannoni C.S., Sherwood M.H., Chuang I.L.: *Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance*. Nature, vol. 414, 2001, pp. 883-887.
- [7] Jacak L.: *Komputer kwantowy: nowe wyzwanie dla nanotechnologii*. Postępy Fizyki, tom dodatkowy 53D, 2002, str. 72-78.
- [8] Vaziri A., Weihs G.: *Experimental Two-Photon, Three-Dimensional Entanglement for Quantum Communication*. Physical Review Letters, vol. 89, 2002, pp. 240401-1-240401-4.
- [9] Biometrics Market and Industry Report 2004-2008. International Biometric Group. 2004.
- [10] Biometrics Boom. IEEE Spectrum, March 2004, p. 9.
- [11] Latest Tests of Biometrics Systems Shows Wide Range of Abilities. IEEE Spectrum Online, Web Only News, <http://www.spectrum.ieee.org/WEBONLY/wonews/jan04/0104biom.html>.
- [12] Materiały reklamowe firmy AutoID Polska - Systemy Automatycznej Identyfikacji Sp. z o.o. Kraków, www.autoid.pl.

OPTOELECTRONIC METHODS FOR PROTECTION OF TELEINFORMATIC INFRASTRUCTURE

Summary

Key importance of teleinformatic infrastructure for the modern-day state necessitates the implementation of adequate protection measures. The paper presents access control methods for sites where elements of this infrastructure are installed. These methods use advanced imaging techniques as well as interferometric fibre optic sensors. Moreover, biometric techniques for user identification and authorization are outlined. Finally, monitoring methods of fibre optic links using optical reflectometry are discussed, as well as novel methods of data transmission which make it possible to detect unauthorised reading of the data transmitted in the link.

Artur Skrygulec, Andrzej Ruciński

Department of Electrical and Computer Engineering,
University of New Hampshire, USA

ZAGADNIENIA NIEZAWODNOŚCI W MIKROSYSTEMACH BEZPIECZEŃSTWA

Streszczenie

W czasach globalnego terroryzmu, mikrosystemy znalazły szerokie zastosowanie w aplikacjach bezpieczeństwa, które łączą w sobie cechy zarówno urządzeń komercyjnych jak i wojskowych. Wymagana jest zarówno ich wysoka niezawodność funkcjonalna jak i odpowiednia trwałość działania. Jednym z przykładów systemów bezpieczeństwa jest przenośny mikrosystem do rozpoznawania linii papilarnych autoryzowanego użytkownika. Problem niezawodności takich systemów jest przedmiotem niniejszej pracy. Koncepcja niezawodności rozszerzona została o aspekty bezpieczeństwa zdefiniowane jako odporność mikrosystemów. Problem odporności mikrosystemów przedstawiony został w aspekcie poboru mocy i temperatury operacyjnej pracy mikrosystemu. Analiza niezawodności mikrosystemów w zależności od temperatury ich pracy jest zaprezentowana na przykładzie wyżej wymienionego mikrosystemu do rozpoznawania linii papilarnych w oparciu o oprogramowanie PRISM służące do szacunków niezawodności.

1. INTRODUCTION

Microsystems¹, for safety and security applications, such as portable and mobile biometric devices, include high functional reliability required for biometrics combined with high durability achieved through ruggedized design. Therefore "traditional" reliability analysis in microsystem [13] design should be expanded. **Reliability** is typically defined as a probability that a microsystem is functioning properly over time interval $<0, t>$ provided that it is functioning at time $t = 0$.

The following figure represents the proposed expansion of design issues with emphasis on reliability improvements in microsystems for biometric applications (described in section 3). FPGA technology has been selected as a SoC implementation.

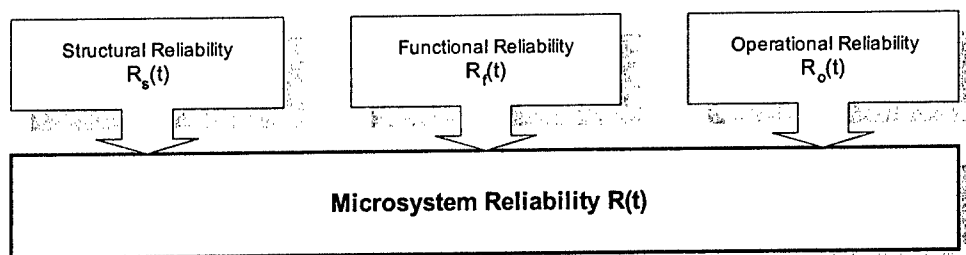


Figure 1. Microsystem Reliability Factors

At least three major tracks have to be distinguished in a process of designing a reliable microsystem for biometrics applications:

- (1) The reliability of a microsystem from the hardware point of view, denoted $R_s(t)$, called **Structural Reliability**, in our case, of biometric fingerprint recognition system [2]. This is a case of reliability enhanced through redundancy ("Structural Reliability" block in Figure 1).
- (2) **Functional Reliability**, $R_f(t)$, relies on the quality of a biometric algorithm influencing to underlying biometric formulas. Functional reliability can be improved, for instance by reconfiguring a SoC architecture ("Functional Reliability" block in Figure 1).
- (3) Issues such as humidity, electromagnetic fields, vibration and others influencing microsystem reliability, are represented in Figure 1 as **Operational Reliability**, $R_o(t)$. One of the most decisive operational factors is temperature. The relationship between temperature and operational reliability is elaborated in Section 3.

In addition to the expanded reliability model depicted in Figure 1, safety and security definitions have to be revised and the relationship among safety, security, and reliability analyzed.

According to the Webster Dictionary [15], safety and security are as follow:

Safety – *in a safe manner; without incurring danger; without hurt or injury; in safety; securely and carefully.* This scenario represents cases where malfunctioning microsystem represents no danger to the rest of the system.

Security – *The state of being secure; freedom from apprehension; confidence of safety sometimes overconfidence; freedom from danger of risk; safety; that which secures or makes safe; surety a person who engages himself for the performance of another's obligations; an evidence of property, as a bond.* In microsystems, this is the ability of a microsystem to withstand a malicious intrusion, e.g. a virus attack, from the outside world, for instance, through encryption.

A continuous Markov Model for unreliable microsystem without repair is depicted in Figure 2.

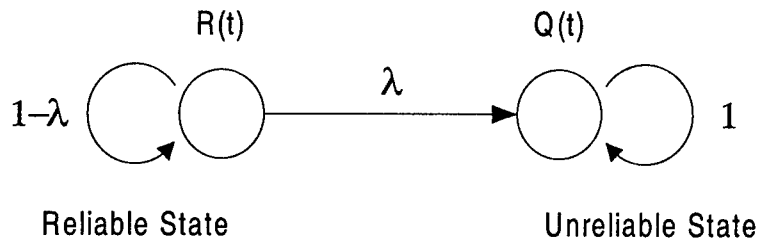


Figure 2. Reliability Model

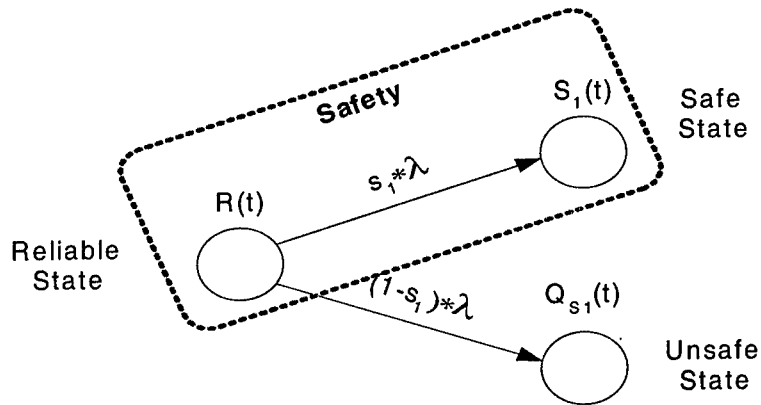
Where:

- $R(t)$ – the probability that a microsystem is functioning properly (reliable state)
- $Q(t)$ – the probability that the microsystem failed (unreliable state)
- λ – failure rate

As stated before, in biometrics applications not only reliability but also safety and security need to be considered. Figure 3a represents a model of a safe system. The failure rate λ is divided into two cases: (1) the failure rate $s_1\lambda$ for transitions from a reliable state into a safe state, and (2) the failure rate for transitions from a reliable state into unsafe state. For example, a microsystem may fail, but a safety mechanism sends a signal generated by self-diagnostics isolating the microsystem and eliminating situations when malfunctioning microsystem may have a negative impact. When the safety coefficient $s_1 = 0$, the safe system (Figure 2a) becomes the reliable system depicted in Figure 1. Figure 2a represents a secure system. In this case, when the system is attacked from the outside, the secure state presents the ability of the system to recognize the attack and to undertake proper precautions. It is assumed that the system under attack is not able to function normally any more since the environment became different. For example, in biometric microsystems, a secure state could be either a detection of false fingerprint or a detection of harmful substances.

The next section briefly introduces a fingerprint authentication microsystem used as a case study. This is followed by the introduction of an extended reliability model that combines reliability, safety, and security. Finally, a design strategy is presented that allows temperature monitoring and hence the dynamic determination of operational reliability of a microsystem.

a.



b.

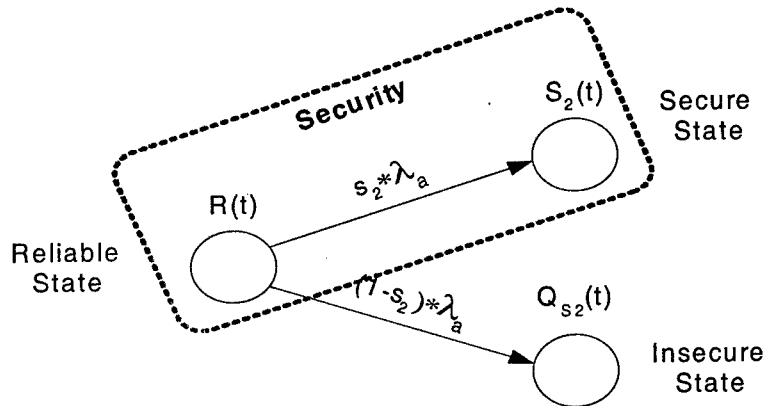


Figure 3. (a) Safety and (b) Security Models

Where:

- $R(t)$ – reliable state
- $S_1(t)$ – safe state
- $S_2(t)$ – secure state
- $Q_{s1}(t)$ – unsafe state
- $Q_{s2}(t)$ – insecure state
- λ – failure rate
- λ_a – attack (intrusion) rate
- s_1 – safety coefficient
- s_2 – security coefficient

2. FINGERPRINT AUTHENTICATION SYSTEM

In this section, a fingerprint authentication microsystem is introduced. The microsystem is based on a patented technology developed by the FPR Corp. [5].

The FPR system is a fast pattern recognizer based on a mixed architecture. It consists of software and digital-analog hardware components. The software part is a PC based application running under Windows NT used for initial processing including image spatial filtering and Fourier transformation. The data processed by the software front end is sent to the hardware – ISA extension card for image matching. One of the unique solutions employed by FPR is using an analogue component on an ISA card. This is a dispersive delay line to increase the operating speed.

The microsystem has originated from the FPR pattern recognizer to obtain fingerprint authentication. The block diagram presented in Figure 4 represents the logical architecture of the developed fingerprint authentication microsystem.

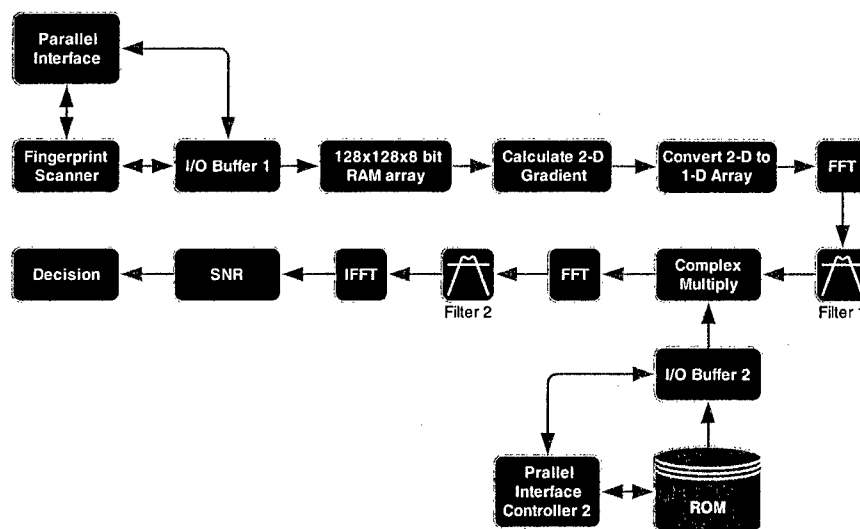


Figure 4. Block Diagram of the Fingerprint Authentication Microsystem Architecture

The unknown image from the fingerprint sensor is sent using a parallel port to the RAM memory located in hardware within the Xilinx Virtex-II [20], [21]. The image is later on processed using two-dimensional filtering by 9×9 coefficients convolution kernel. The gradient obtained in such a way is reorganized and forms one dimensional array of complex points that in turn are processed by the FFT core [3]. Filter 1 represents cutting off negative frequency components from the spatial-frequency data received from FFT. Modified in such a way spatial-frequency data represents the unknown pattern which is correlated against the templates stored in ROM. Each template is complex multiplied with the unknown pattern and processed by the digital equivalent of the dispersive delay line depicted by the FFT – Filter 2 – IFFT set of blocks. The output of a dispersive delay line forms a signal that may contain significant amplitude spikes. These spikes, if present, correspond to a match for a given template. Based on adopted fingerprint authentication architecture the MatLab [17] prototype was developed as illustrated in Figure 5.

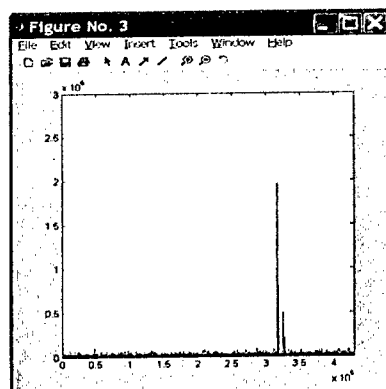


Figure 5. MatLab Model of the Fingerprint Authentication System

The virtual prototype developed in Matlab allowed for the system architecture characterization and debugging before hardware implementation. The Matlab model has shown 0% erroneously accepted users, while around 94% of authorized users have been recognized correctly.

Next, the biometric microsystem was developed in VHDL and adjusted for the following configuration: Virtex-II XC2V8000 FPGA from Xilinx [20], [21], AES3500 capacitive fingerprint sensor from Authentec [2]. The fingerprint sensor produces 128 x 128 8 byte gray scale bitmap, which is sent to the FPGA via standard parallel port communication interface. The transmission speed allows sending the entire image to the hardware processing core in a little less than 200 ms.

When the image is received and stored in the internal RAM memory located within FPGA the recognition process is initiated. It completes within 20 ms using 45 MHz clock. It has been determined that a MatLab [17] simulation model running on a Dell PC with Pentium4 2.4 GHz processor required 2.14 sec for a single fingerprint recognition. This means that the hardware version is approximately 100 times faster than its PC based implementation and further significant improvements are possible.

The following Table 1 provides the summary of resources used to implement fingerprint authentication system in the FPGA Xilinx Virtex-II technology.

Table 2.1

Virtex-II Resource Usage

Resource type	Total Amount	Amount Used	Usage Percentage
Number of Slices	46,592	9,809	21%
Total Number Slice Registers	93,184	16,271	17%
Total Number 4 input LUTs	93,184	11,693	12%
Number of bonded IOBs	824	16	1%
Number of Tbufs	23,296	64	1%
Number of Block RAMs	168	88	52%
Number of MULT18X18s	168	41	24%
Number of GCLKs	16	4	25%
Total equivalent gate count for design		6,483,959	

3. ROBUSTNESS MODEL FOR BIOMETRIC MICROSYSTEMS

The expanded reliability model is described in this section. It considers all reliability components presented previously: reliability, safety, and security. The reliability part includes all tracks introduced in Section 1: *structural*, *functional*, and *operational*. In order to integrate all the above components, the robustness of a microsystem is defined as follows:

Robustness $\mathcal{R}(t)$ is the probability that the microsystem fulfilling the following conditions:

- (i) it is functioning properly over time interval $\langle 0, t \rangle$, or
- (ii) it is in the safe state in case of malfunctioning over time interval $\langle 0, t \rangle$, or
- (iii) it is in the secure state in case of either external attack or emergency.

It is assumed that the microsystem is functioning properly at time $t=0$. The corresponding Markov based continuous model without repair is presented in Figure 6. For simplicity, it is assumed that **failures** and **external attacks** are independent events and that do not occur simultaneously, e.g. the probability of the failure and an attack at a time is negligible. In reality, if an attack causes a failure, the system is malfunctioning and it is in unsafe state.

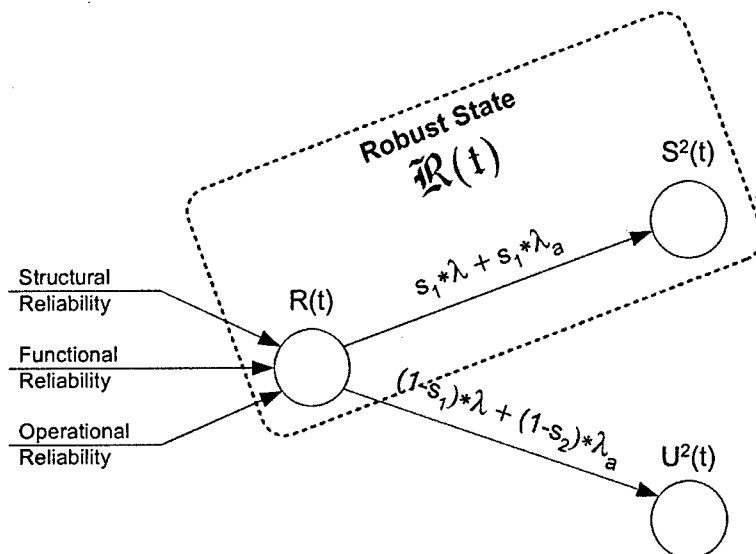


Figure 6. Microsystem Robustness Model

Where:

- $R(t)$ – reliable state
- $S^2(t)$ – safe and secure state
- $U^2(t)$ – unsafe and insecure state
- λ – failure rate
- λ_a – attack (intrusion) rate
- s_1 – safety coefficient
- s_2 – security coefficient

Note, that the probability that the microsystem is in safe and secure state $S^2(t)$ is greater than the probability that the system is in safe state $S_1(t)$ (Figure 3a) due to security measures. Again, in particular systems security such as cryptography and safety like diagnostics and health monitoring may overlap. Safety and security coefficients, i.e. s_1 and s_2 respectively, represent conditional probabilities provided either a failure or an attack takes place. It has to be noted that the proposed robustness model is one of many others possible for consideration, analysis, and validation impractical settings. The remaining part of this section provides more detailed analysis of different reliability components and relationships among them.

3.1 Structural Reliability Assessment

Structural reliability $R_s(t)$ can be improved by introducing redundant components. For example, a Triple Modular Redundancy (TMR) arrangements assumes that the basic module with reliability $R(t)$ is triplicated and the resulting output is generated by a majority vote as depicted in Figure 7.

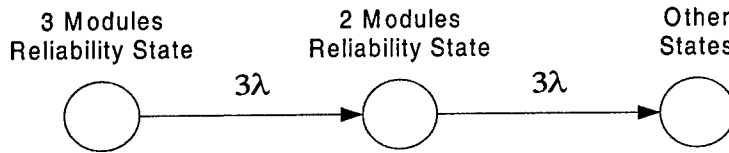


Figure 7. TMR Model

Assuming that only one module can fail at a time and that all module reliabilities are equal and described by exponential law, we have:

$$R_{TMR}(t) = R^3(t) + 3R^2(t) \cdot (1 - R(t)) \quad (3.1.1)$$

$$R_{TMR}(t) \geq R(t) \quad (3.1.2)$$

$$\text{if } t < \frac{1}{\lambda} \quad (3.1.3)$$

3.2 Functional Reliability Enhancement

Functional reliability $R_f(t)$ in biometrics is derivative of hardware reliability. The function of a biometric system is to properly recognize the object, determined by the False Acceptance Rate (FAR), and properly reject the object, determined by the False Rejection Rate (FRR). However, FRRs and FARs depend on different biometric algorithms. This drawback is one of the reasons why a universal biometric system does not exist yet. In addition, a microsystem implemented in FPGA can be restructured and hence different FRRAs and FARs can be obtained increasing the microsystem reliability. At the same time, as a rule, this enhancement would require more hardware resulting in decreased reliability. This scenario is illustrated in Figure 8.

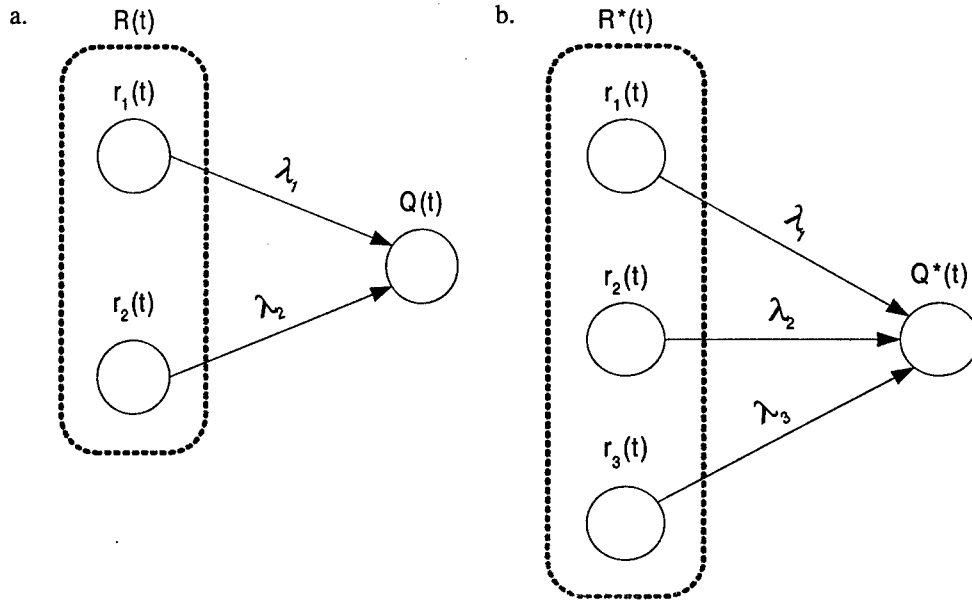


Figure 8. Functional Reliability: (a) Example of Less Robust System,
(b) Example of System with Improved Robustness

Hence, we have

$$R_f^*(t) \geq R_f(t) \text{ and} \quad (3.2.1)$$

$$R_s^*(t) \leq R_s(t) \quad (3.2.2)$$

with some robustness improvement $\Delta \mathcal{R}(t)$ is still undetermined.

3.3. Operational Reliability Evaluation

In this section the elaboration on operational reliability is provided. Since power management and temperature controlling can improve reliability, it is possible that lowering power consumption leads to lower operating temperature of the microsystem. A relevant design methodology in microsystems is summarized in Figure 9.

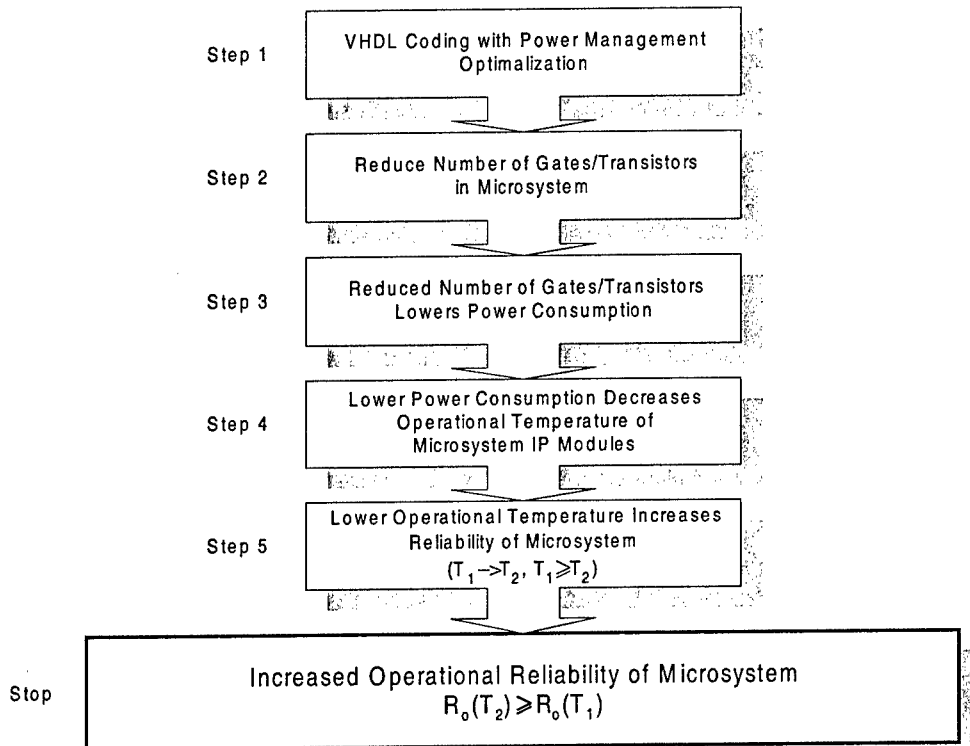


Figure 9. Operational Reliability Improvement Model

An overview of components of the proposed methodology is as following:

- **Step 1** [Figure 9] – any method for VHDL algorithm optimization such as the Adaptive Evolutionary Algorithm [7]. Decreasing the number of gates using in the microsystem and hence decreasing the number of transistors leads to lower power consumption. The concept is based on identification of non critical paths modules in data flow and lowering the operational clock frequency in those modules.
- **Step 2** – one can find a relation between the number of gates/transistors and power consumption of a microsystem. For instance, let us consider a 0.5-mm technology CMOS device with 500,000 gates clocked at 100 MHz. If each gate dissipates 1 $\mu\text{W}/\text{MHz}$ and if 25% of the gates are toggling at a time, the chip dissipates 12.5W of power [8]. Since the number of gates is directly proportional to the power consumption the next relationship holds.
- **Step 3** – typically, either a microsystem's data sheet or packaging guide provides specifications for determining a device's maximum power consumption. These specifications include [4]:
 - T_J , or maximum allowed junction temperature of the silicon in degrees Celsius,
 - T_A , or allowed ambient temperature range in degrees Celsius, and

- *uppercase theta*_{JA}, or junction-to-ambient thermal resistance of the device/package combination in degrees Celsius per watt. After determining system's worst-case ambient temperature, the maximum allowable device power consumption in the selected package is

$$P_{MAX} = \frac{T_J - T_A}{\text{uppercase theta}_{JA}} \quad (3.3.1)$$

Where:

P_{MAX} – is the maximum allowable device power consumption.

According to the above equation, the temperature is directly proportional to the power consumption. Hence, lowering power consumption decrease the working temperature of a chip.

- **Step 4** – by decreasing the operating temperature device improves the reliability of the microsystem. Using data provided by RAC Laboratories [9], [16], the microsystem reliability approximations have been evaluated using PRISM modeling package [6], [10]. Figures 10 and 11 represent estimated behavior of the biometric fingerprint recognition system introduced in Section 2 versus the operating and dormant temperature respectively.

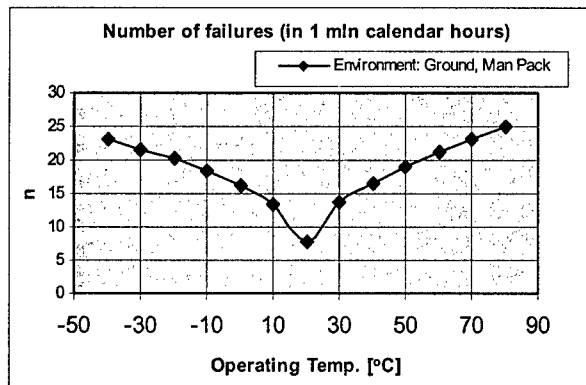


Figure 10. Number of Failures (n) Versus Operating Temperature T_O (°C)

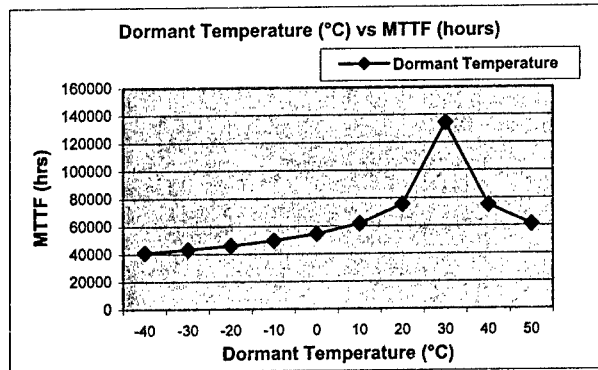


Figure 11. MTTF¹ (hrs) Versus Dormant Temperature T_D (°C)

- **Step 5** – at this point, one can state that power management and thermal monitoring is the key challenge for operational reliability enhancement in the microsystem.

Properties of robustness can be derived by analyzing relationships among its different components. Let us consider a biometric microsystem with \mathcal{R} , R_s , R_f , R_o , s_1 , and s_2 probabilities. In general, the robustness can be improved if either R_s or R_f or R_o varies. Also, to have meaningful microsystem, $R_s, R_f, R_o \gg s_1, s_2$. We have

$$\mathcal{R} = \min\{R_s, R_f, R_o\} \quad (3.3.2)$$

$$R_o \geq R_s \geq R_f \geq s_1 \geq s_2 \quad (3.3.3)$$

$$R_f \in |FAR - FRR| \quad (3.3.4)$$

$$R_o(t) \approx |T_o - 20| + 5 \quad (3.3.5)^2$$

All reliabilities are correlated and any improvement of any of them results in improvement of the system robustness.

4. CONCLUSIONS

A formal model of robust microsystems for biometric applications was introduced. It integrates the notions of reliability, safety, and security. Two innovative ways are shown related to the robustness of such microsystems: 1) Reconfiguration of a microsystem that increases its functional reliability; and 2) Using operational temperature as a parameter determines the operational reliability and hence the microsystem robustness.

¹ RAC PRISM generates MTTF (Mean Time To Failure) in Failures/ Million Hours

² Approximation based on Figure 10

ACKNOWLEDGEMENTS

The authors acknowledge help and contribute of the following individuals and organizations: Dr. John Apostolos, FPR Corporation, Dr. Henry Mullaney, Director of the New Hampshire Industrial Research Center, Dr. Władysław Szczepniak, Gdańsk University of Technology, and Xilinx Inc.

BIBLIOGRAFIA

- [1] Altet J. and Rubio A.: *Thermal Testing of Integrated Circuits*, Dordrecht, The Netherlands, 2002.
- [2] AuthenTec Inc.: *EntrePad AES3500 Fingerprint Sensor Product Family Specification*, www.authentec.com.
- [3] Dillon Engineering Inc.: *Ultra High-Performance FFT/IFFT IP Core*, www.dilloneng.com
- [4] Dipert B., ed.: *Programmable logic: Beat the heat on power consumption*, EDN Access for Design by Design, August 1, 1997, http://www.e-insite.net/ednmag/archives/1997/080197/16df_01.htm#Measuring%20temperature%20the%20all-natural%20way
- [5] FPR Corporation: *Fast Pattern Recognizer Utilizing Dispersive Delay Line*, United States Patent no. 5,859,930.
- [6] <http://rac.alionscience.com/prism/>
- [7] Kozieł S. and Szczepniak W.: *High Level Synthesis with Adaptive Evolutionary Algorithm for Solving Reliability and Thermal Problems in Reconfigurable Microelectronic Systems*, September 2003, Aix-en-Provence, France.
- [8] Lipman J., ed.: *EDA tools let you track and control CMOS power dissipation*, EDN Access for Design by Design, November 23, 1995, <http://www.reed-electronics.com/ednmag/archives/1995/112395/24df3.htm#fig2>
- [9] rac.alionscience.com
- [10] Reliability Analysis Center: *PRISM User Manual. V 1.2*, Rome, NY 2002
- [11] Richards M.A. et. al.: *Rapid Prototyping of Application Specific Signal Processors*, Kluwer Academic Publishers, Boston, 1997.
- [12] Ruciński A., Skrygulec A., and Mocny J.: *A VIRTEX-II Based Biometric Microsystem for Fingerprint Authentication*, Gdańsk, Poland, May 2003.
- [13] Ruciński A., Skrygulec A., Pysareva K., and Mocny J.: *Microsystem Development Using the TQM Design Methodology*, Perth, Australia, January 2004.
- [14] Smith Winthrop W., Smith Joanne M.: *Handbook of Real-Time Fast Fourier Transforms. Algorithms to Product Testing*, IEEE, 1995.
- [15] Thatcher V.S., ed.: *The New Webster Encyclopedic Dictionary of the English Language*, Evenel Bodes, N.Y., 1980
- [16] www.eda.org/rassp/index.html
- [17] www.mathworks.com
- [18] www.mentor.com
- [19] www.nhirc.unh.edu
- [20] Xilinx Inc.: *Coregen Intellectual Property Library Technical Documentation*, www.xilinx.com.
- [21] Xilinx Inc.: *Virtex-II Field Programmable Gate Arrays Technical Documentation*, www.xilinx.com.

ZAGADNIENIA NIEZAWODNOŚCI W MIKROSYSTEMACH BEZPIECZEŃSTWA

Streszczenie

W czasach globalnego terroryzmu, mikrosystemy znalazły szerokie zastosowanie w aplikacjach bezpieczeństwa, które łączą w sobie cechy zarówno urządzeń komercyjnych jak i wojskowych. Wymagana jest zarówno ich wysoka niezawodność funkcjonalna jak i odpowiednia trwałość działania. Jednym z przykładów systemów bezpieczeństwa jest przenośny mikrosystem do rozpoznawania linii papilarnych autoryzowanego użytkownika. Problem niezawodności takich systemów jest przedmiotem niniejszej pracy. Koncepcja niezawodności rozszerzona została o aspekty bezpieczeństwa zdefiniowane jako **odporność** mikrosystemów. Problem **odporności** mikrosystemów przedstawiony został w aspekcie poboru mocy i temperatury operacyjnej pracy mikrosystemu. Analiza niezawodności mikrosystemów w zależności od temperatury ich pracy jest zaprezentowana na przykładzie wyżej wymienionego mikrosystemu do rozpoznawania linii papilarnych w oparciu o oprogramowanie PRISM służące do szacunków niezawodności.

Marek Blok

Katedra Systemów Informacyjnych, Politechnika Gdańska

KONWERSJA SYGNAŁÓW CYFROWYCH POMIĘDZY TELEKOMUNIKACYJNYMI I MULTIMEDIALNYMI SZYBKOŚCIAMI PRÓBKOWANIA

Streszczenie

W pracy omówiono problematykę projektowania algorytmów konwersji szybkości próbkowania dla potrzeb realizacji styku pomiędzy systemami pracującymi z telekomunikacyjnymi oraz multimedialnymi szybkościami próbkowania. Omówiono tutaj klasyczne podejście do projektowania systemów przepróbkowania wskazując na jego ograniczenia w aspekcie rozpatrywanego zastosowania. Jako alternatywne rozwiązanie w pracy zaproponowano algorytm konwersji realizowany na bazie filtrów ułamkowo-opóźniających optymalnych w sensie Czebyszewa. Skuteczność tego rozwiązania zaprezentowano na przykładach projektów algorytmu, w tym rozwiązania dwustopniowego znacząco obniżającego wymagania pamięciowe implementacji, przy założeniu szczególnie wysokiej jakości realizowanej konwersji.

1. WSTĘP

Współcześnie częstym problem na styku dwóch systemów wymiany informacji audiowizualnej jest niezgodność standardów szybkości próbkowania, zwłaszcza gdy stosunek szybkości próbkowania jest w znacznym stopniu niewspółmierny. Przykładem może być tutaj połączenie systemów przekazujących sygnały akustyczne, pracujących z szybkościami typowymi dla urządzeń telekomunikacyjnych (wielokrotności 8 000 próbek/s) z systemami pracującymi z szybkościami multimedialnymi (wielokrotności 11 025 próbek/s). Wymiana sygnałów dyskretnych pomiędzy takimi systemami wymaga konwersji szybkości próbkowania przekazywanego sygnału. Najprostszym rozwiązaniem jest konwersja sygnału z postaci cyfrowej na analogową, a następnie ponowne próbkowanie z nową szybkością. Oznacza to jednak nieuchronne pogorszenie jakości sygnału. Alternatywnym rozwiązaniem jest wykonanie konwersji całkowicie po stronie cyfrowej. Niestety w przypadku przepróbkowania sygnałów pomiędzy standardami telekomunikacyjnymi i multimedialnymi, ze względu na wysoką niewspółmierność stosowanych tam szybkości próbkowania, nie można zastosować klasycznych algorytmów przepróbkowania. Dlatego też w pracy zaproponowano zastosowanie algorytmu wykorzystującego ułamkowo-opóźniające filtry FIR optymalne w sensie Czebyszewa. Działanie algorytmu zilustrowano na przykładzie konwersji pomiędzy szybkościami: 48 000 próbek/s i 44 100 próbek/s.

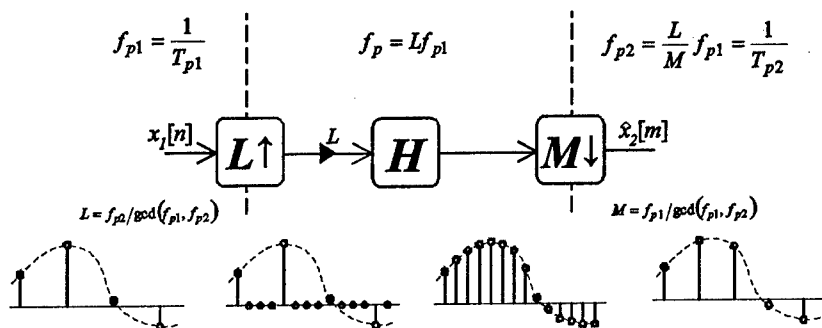
2. SFORMUŁOWANIE PROBLEMU

Klasyczne i jednocześnie najczęściej stosowane podejście do zmiany szybkości próbkowania sygnałów przedstawia rys. 1 [1, 2]. W pierwszym etapie (blok $L\uparrow$) szybkość próbkowania sygnału wejściowego $x_1[n]$ jest zwiększana L -krotnie poprzez wstawienie pomiędzy każde dwie sąsiednie próbki $L-1$ zer (ang. *zeroinserting*), co na skali unormowanej częstotliwości skutkuje kompresją, a zarazem replikacją widma sygnału wejściowego. Jednocześnie, aby poziom każdej z replik widma był taki sam jak poziom widma sygnału oryginalnego, konieczne jest L -krotne wzmocnienie sygnału. W drugim etapie filtr interpolacyjno-decymacyjny, czyli dolnoprzepustowy filtr cyfrowy H z rys. 1 o unormowanej częstotliwości granicznej

$$v_g = \min(0.5/L, 0.5/M) \quad (2.1)$$

oblicza wartości interpolowanych próbek, usuwając zbędne repliki widmowe. Następny blok $M\downarrow$, kompresor skali czasu, pozostawia co M -tą próbkę zinterpolowanego sygnału pomijając pozostałe; skutkuje to ekspansją widma sygnału. Dla $L < M$, filtr interpolacyjno-decymacyjny poza usuwaniem replik widma zapobiega efektowi aliasingu (nakładania się na siebie „ogonów” replik widma).

Wadą bezpośredniej implementacji algorytmu przepróbkowania z rys. 1 jest jego bardzo niska efektywność numeryczna. Jak można zauważyć znaczna część próbek na wejściu filtru interpolacyjno-decymacyjnego H jest zerowa, ponadto tylko część z próbek wyjściowych tego filtru jest ostatecznie pozostawiana.



Rys.1. Klasyczny algorytm zmiany szybkości próbkowania

Rozpatrzmy przykładowo konwersję z szybkości $f_{p1} = 48\,000$ próbek/s na szybkość $f_{p2} = 44\,100$ próbek/s. Całkowite współczynniki L i M zależą od wejściowej i wyjściowej szybkości próbkowania

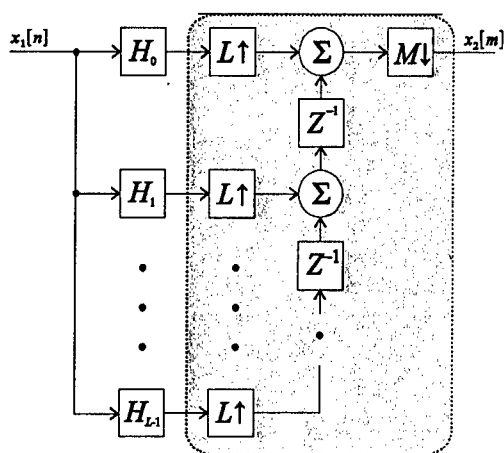
$$L = f_{p2} / \gcd(f_{p1}, f_{p2}) \quad \text{oraz} \quad M = f_{p1} / \gcd(f_{p1}, f_{p2}) \quad (2.2)$$

gdzie $\gcd(x, y)$ oznacza największy wspólny dzielnik liczb x i y . W naszym przypadku $L=147$ i $M=160$. Stąd też przejściowa szybkość próbkowania z jaką powinien pracować filtr interpolacyjno-decymacyjny z rys. 1 wynosi aż $7\,056\,000$ próbek/s. Jednocześnie zauważmy, że wymagana unormowana częstotliwość graniczna filtru interpolacyjno-decymacyjnego jest bardzo mała, co oznacza, że aby zapewnić dobrą jakość przepróbkowania odpowiedź impulsowa tego filtru musi być bardzo długa (nawet rzędu kilku tysięcy próbek). Tak duża przejściowa szybkość próbkowania w połączeniu z bardzo długą odpowiedzią impulsową filtru interpolacyjno-decymacyjnego oznacza olbrzymie koszty

numeryczne, w praktyce nie do zaakceptowania. Problem ten jednak można stosunkowo prosto rozwiązać poprzez odpowiednią reorganizację algorytmu, pomijając te obliczenia, które są zbędne, czyli dotyczące próbek zerowych lub próbek, których ostatecznie się nie wykorzystuje. Efektem takich optymalizacji są struktury polifazowe [1, 2]. Schemat przykładowej implementacji algorytmu przepróbkowania wykorzystujący strukturę polifazową przedstawia rys. 2. Odpowiedź impulsowa $h_p[n]$ p -tego polifazowego filtra H_p otrzymuje się na podstawie odpowiedzi impulsowej filtra interpolacyjno-decymacyjnego $h_l[n]$ z klasycznego algorytmu poprzez jej polifazową dekompozycję

$$h_p[n] = h_l[p + nL]; \quad p = 0, 1, \dots, L-1 \quad (2.3)$$

Wszystkie obliczenia są tu wykonywane z wejściową szybkością próbkowania a, zadaniem części algorytmu wyróżnionej szarym tłem, jest określenie na wyjściu którego filtra polifazowego znajduje się bieżąca próbka wyjściowa algorytmu przepróbkowania. Efektem zastosowania dekompozycji polifazowej filtra interpolacyjno-decymacyjnego jest L -krotne zmniejszenie złożoności numerycznej algorytmu w stosunku do rozwiązania klasycznego. Dodatkowe M -krotne zmniejszenie złożoności numerycznej można uzyskać nie obliczając wartości próbek filtrów polifazowych, gdy nie są one ostatecznie potrzebne.



Rys.2. Schemat blokowy implementacji klasycznego algorytmu zmiany szybkości próbkowania w stosunku L/M wykorzystującej strukturę polifazową

Bardziej istotna wada klasycznego podejścia do przepróbkowania, dotycząca również rozwiązań polifazowych, wiąże się z koniecznością projektowania wąskopasmowego filtra dolnoprzepustowego. Zadanie to nie nastręcza trudności, gdy krotności: interpolacji L i decymacji M , są rzędu kilku lub kilkunastu. Jednak dla większych wartości, w połączeniu z wysokimi wymaganiami na jakość konwersji, zaprojektowanie filtra stanowi bardzo poważne wyzwanie, ze względu zarówno na bardzo wąskie pasma: przepustowe oraz przejściowe, jak i bardzo długą odpowiedź impulsową tego filtra. Przykładowo, dla interesującej nas tutaj konwersji z 48 000 próbek/s na 44 100 próbek/s, przy założeniu, że poniżej częstotliwości 19 kHz odchylenie charakterystyki amplitudowej nie będzie większe niż 0.1dB, a tłumienie składowych mogących spowodować zniekształcenia aliasingowe będzie nie mniejsze niż 100dB, szacowana długość odpowiedzi impulsowej filtra interpolacyjno-decymacyjnego optymalnego w sensie kryterium MINIMAX [3] wynosi aż 9 413 próbek. Należy jednak pamiętać, że zaprojektowanie takiego filtra optymalnego, ze

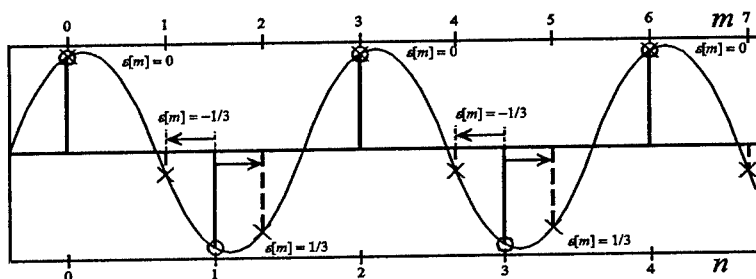
względem na problemy numeryczne, związane przede wszystkim z koniecznością odwracania olbrzymich macierzy, jest obecnie nieosiągalne. Z kolei zastosowanie rozwiązań nieoptymalnych znacznie wydłuża odpowiedź impulsową, a jednocześnie utrudnia osiągnięcie rozwiązania spełniającego narzucone wymagania.

Jednym z rozwiązań powyższego problemu jest implementacja zmiany szybkości próbkowania w zaokrąglonym stosunku szybkości próbkowania [4]. Przykładowo, zamiast $48000:44100 = 160:147$, można zastosować przepróbkowanie $48:44 = 12:11$. Rozwiązanie takie znacznie ułatwia zadanie projektowania algorytmu przepróbkowania, jednak ma ono ograniczone zastosowania. Można je użyć, gdy przetwarzaniu poddaje się wyłącznie fragmenty sygnałów i obserwowana nieznaczna kompresja albo ekspansja czasu nie jest istotna lub w zastosowaniach, gdy odchyłka wynikowej szybkości próbkowania od nominalnej jest do przyjęcia i nie wpływa na pracę całego systemu. Nie można jednak zastosować tego rozwiązania przy łączeniu niezależnych systemów pracujących w czasie rzeczywistym, gdzie, dla powyższego przykładu, na wyjściu algorytmu przepróbkowania co sekundę powstaje niedobór 100 próbek.

3. ALGORYTM WYKORZYSTUJĄCY FILTRY OPÓŹNIAJĄCE

Stosunek szybkości próbkowania sygnałów telekomunikacyjnych do szybkości próbkowania sygnałów multimedialnych jest na tyle niekorzystny, że w praktyce wyklucza zastosowanie klasycznego podejścia z rys. 1. W takiej sytuacji stosuje się rozwiązanie realizowane w oparciu o filtry ułamkowo-opóźniające [5-9].

Zauważmy, że aby wyznaczyć próbkę wyjściową $x_2[m]$, należy określić wartość sygnału $x(t)$, reprezentowanego przez ciąg próbek wejściowych $x_1[n]$, w nowych chwilach czasu (rys. 3).



Rys.3. Ilustracja 3/2-krotnego przepróbkowania sygnałów przy założeniu $T_{p1} = 1$; o – próbki sygnału przed przepróbkowaniem $x_1[n]$ oraz x – próbki sygnału po przepróbkowaniu $x_2[m]$.

Zapiszmy teraz ciąg próbek wyjściowych $x_2[m]$ w odniesieniu do okresu próbkowania ciągu wejściowego $x_1[n]$

$$x_2[m] \triangleq x(mT_{p2}) = x(n_n[m] - \varepsilon[m]T_{p1}); \quad \forall m, n \in I \quad (3.1)$$

gdzie T_{p1} jest okresem próbkowania ciągu $x_1[n]$, T_{p2} jest okresem próbkowania ciągu $x_2[m]$ a $n_n[m]$ jest liczbą całkowitą taką, że $\varepsilon[m] \in [-1/2, 1/2)$. Wówczas zagadnienie konwersji szybkości próbkowania możemy sformułować następująco: na podstawie ciągu $x_1[n]$ oszacować wartości tego ciągu dla chwil $n_n[m] - \varepsilon[m]$, tworząc ciąg $x_2[m]$:

$$x_2[m] = \hat{x}_1(n_n[m] - \varepsilon[m]) \quad (3.2)$$

gdzie $n_n[m]$ oznacza numer próbki ciągu $x_1[n]$ najbliższej próbce $x_2[m]$, zaś $\varepsilon[m]$ określa opóźnienie próbki $x_2[m]$ względem próbki $x_1[n_n[m]]$. Powyższy zapis należy interpretować następująco: wartość próbki $x_2[m]$ ciągu wyjściowego jest ułamkowo opóźnioną (o $\varepsilon[m]$) próbką ciągu $x_1[n]$ o numerze $n = n_n[m]$. Przy tym, tylko wtedy, gdy $\varepsilon[m] = 0$, powyższy zapis jest ścisły i $x_2[m] = x_1[n_n[m]]$. Dla pozostałych wartości $\varepsilon[m]$ próbkę wyjściową trzeba oszacować na podstawie próbek ciągu wejściowego $x_1[n]$ położonych wokół próbki o numerze $n_n[m]$, np. za pomocą filtru ułamkowo-opóźniającego. Zatem, żeby wyznaczyć ciąg $x_2[m]$ potrzebna jest znajomość ciągów $n_n[m]$ oraz $\varepsilon[m]$.

Wychodząc z interpretacji tych ciągów możemy zapisać

$$n_n[m] \triangleq \text{round}(mT_{p2}/T_{p1}) = \text{round}\left(m \frac{M}{L}\right) \in I \quad (3.3)$$

gdzie $\text{round}(\cdot)$ oznacza zaokrąglenie do najbliższej liczby całkowitej, oraz

$$\varepsilon[m] \triangleq (n_n[m]T_{p1} - mT_{p2})/T_{p1} = n_n[m] - m \frac{M}{L} \in \langle -1/2, 1/2 \rangle \quad (3.4)$$

gdzie parametry M i L (2.2) mają taką samą interpretację jak w koncepcji klasycznej z rys. 1.

Jednak, ponieważ ciąg $n_n[m]$ narasta do nieskończoności, w praktycznym zastosowaniu przydatniejsze są zależności pozwalające wyznaczać iteracyjnie [10,11] opóźnienie $\varepsilon[m]$:

$$\varepsilon[m] = \text{round}(\varepsilon[m-1] + \frac{M}{L}) - (\varepsilon[m-1] + \frac{M}{L}) \quad (3.5)$$

oraz przyrosty parametru $n_n[m]$:

$$\Delta n_n[m] = n_n[m] - n_n[m-1] = \frac{M}{L} + \varepsilon[m] - \varepsilon[m-1] \in I \quad (3.6)$$

Łatwo można wykazać, że ciąg $\varepsilon[m]$ jest okresowy z okresem L , a co za tym idzie i przyrosty $\Delta n_n[m]$ są również ciągiem okresowym z okresem L . Cecha ta jest bardzo ważna w implementacji. Oznacza to, że aby zaimplementować omawiany algorytm wystarczy rozpatrzyć jedynie L filtrów ułamkowo-opóźniających. Filtry te można zaprojektować starannie zawczasu i później realizując przepróbkowanie pobierać ich współczynniki z tablicy LUT (ang. *look-up table*). Pozwala nam to na zastosowanie do projektowania filtrów ułamkowo-opóźniających kosztownych numerycznie metod projektowania filtrów optymalnych bez podnoszenia kosztów implementacji algorytmu przepróbkowania. Dodatkowo, krotność decymacji M wpływa tylko na kolejność filtrów ułamkowo-opóźniających i przy jednakowym L a różnych M , można wykorzystać te same filtry.

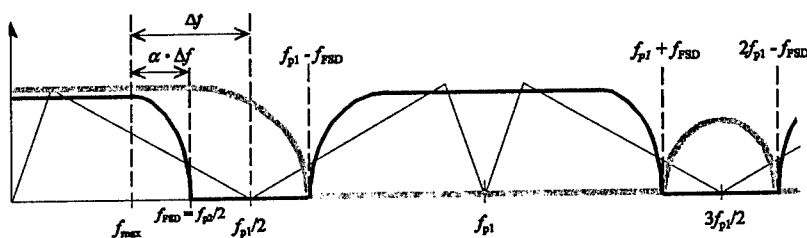
Warto również zauważyć, że algorytm przepróbkowania wykorzystujący filtry ułamkowo-opóźniające można sprowadzić do rozwiązania z rys. 2, z tą tylko różnicą, że zamiast filtrów uzyskanych z dekompozycji polifazowej (2.3) stosujemy filtry ułamkowo-opóźniające

$$h_p[n] = h_{\varepsilon[p]}[n]; \quad p = 0, 1, \dots, L-1 \quad (3.7)$$

gdzie $h_{\varepsilon[p]}[n]$ jest odpowiedzią impulsową filtru ułamkowo-opóźniającego o opóźnieniu $\varepsilon[p]$. Jednak w odróżnieniu od klasycznego algorytmu wymagającego zaprojektowania filtru o bardzo wąskim pasmie i bardzo długiej odpowiedzi impulsowej, w rozwiązaniu tym nawet przy niekorzystnych stosunkach szybkości próbkowania, wymagane jest zaprojektowanie wielu filtrów o krótkich odpowiedziach impulsowych, nie stwarzających większych problemów przy ich projektowaniu. Dodatkowo spostrzeżenie to pozwala na

łatwą całościową ocenę jakości algorytmu przepróbkowania wykorzystującego wiele filtrów o różnych parametrach, w oparciu o koncepcję filtru zbiorczego [12] – odpowiednika filtru interpolacyjno-decymacyjnego.

Niestety podejście wykorzystujące filtry ułamkowo opóźniające nie jest również pozbawione wad. Zastosowanie wielu odrębnie projektowanych filtrów w miejsce jednego filtru interpolacyjno-decymacyjnego skutkuje pojawieniem się artefaktów w pasmie zaporowym filtru zbiorczego [13]. Artefakty te obserwujemy jako znaczne pogarszanie się tłumienia filtru zbiorczego wokół częstotliwości $(2k+1)/2 \cdot f_{p1}$, gdzie $k = 1, 2, \dots$, co schematycznie pokazano na rys. 4. W efekcie zjawisko to może być przyczyną powstawania zniekształceń aliasingowych i należy je uwzględnić w procesie projektowania algorytmu przepróbkowania. Położenie obserwowanych artefaktów jest jednak na tyle korzystne, że zniekształceniom aliasingowym można zapobiec ograniczając pasmo przetwarzanego sygnału (rys. 4). Można to uzyskać stosując przed przepróbkowaniem dodatkową filtrację dolnoprzepustową.



Rys. 4. Ilustracja założeń projektowych algorytmu przepróbkowania ($L=3$ i $M=4$).

Gruba szara linia ilustruje charakterystykę filtru zbiorczego, gruba czarna linia – charakterystykę częstotliwościową wstępnego filtru dolnoprzepustowego, a trójkąty wykreślone cienką linią ilustrują widmo sygnału poddawanego przepróbkowaniu, wraz z jego replikami.

Do tej pory pokazaliśmy, że algorytm przepróbkowania wykorzystujący filtry ułamkowo-opóźniające można skutecznie zastosować do implementacji algorytmu przepróbkowania pomiędzy szybkościami telekomunikacyjnymi i multimedialnymi. Pozostaje jednak problem skutecznego sposobu przeniesienia założeń projektowych algorytmu przepróbkowania na wymagania projektowe dotyczące wstępnego filtru dolnoprzepustowego oraz poszczególnych filtrów ułamkowo-opóźniających. Wybór do tego zastosowania filtrów ułamkowo-opóźniających optymalnych w sensie Czebyszewa pozwala problem ten sprowadzać do kilku prostych reguł [13].

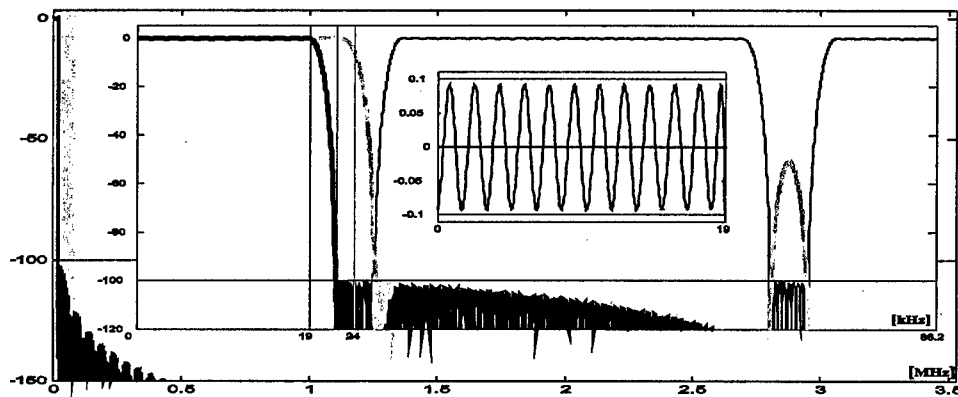
Założenia projektowe algorytmu zmiany szybkości próbkowania z f_{p1} na f_{p2} są podane w postaci następujących parametrów: f_{\max} – częstotliwość poniżej której składowe sygnału powinny być przenoszone przy jak najmniejszych zniekształceniach, A_p – maksymalne dopuszczalne zafalowanie charakterystyki algorytmu przepróbkowania poniżej częstotliwości f_{\max} oraz A_s – minimalne tłumienie replik widma podstawowego. Przy tych założeniach specyfikacja wstępnego filtru dolnoprzepustowego jest postaci: tłumienie w pasmie zaporowym A_s , maksymalne dopuszczalne zafalowanie w pasmie przepustowym A_p , pasmo zaporowe w przedziale od f_{\max} do $f_{\max} + \alpha \Delta f$, gdzie $\Delta f = f_{p1}/2 - f_{\max}$, a α jest arbitralnie dobranym parametrem z przedziału $(0, 1)$. Z kolei specyfikacja filtrów ułamkowo-opóźniających optymalnych w sensie Czebyszewa jest postaci: częstotliwość graniczna pasma aproksymacji $f_{\text{FSD}} = f_{\max} + \alpha \Delta f$, a dopuszczalny błąd w paśmie

aproksymacji wynosi $-A_s$. Należy jeszcze przy projektowaniu filtrów pamiętać, że szybkość próbkowania z jaką pracują rozpatrywane filtry, to wejściowa szybkość próbkowania f_{p1} .

Jedynym otwartym problemem w podanym powyżej „przepisie” na projektowanie algorytmu zmiany szybkości próbkowania jest dobór parametru α . Parametr ten odpowiada bezpośrednio za wymiennność pomiędzy szerokością pasma przejściowego wstępnego filtru dolnoprzepustowego, a szerokością pasma niespecyfikowanego filtrów ułamkowo-opóźniających. Przenosi się to pośrednio na wymiennność pomiędzy długościami odpowiedzi impulsowych tych filtrów. Ze wzrostem wartości parametru α pasmo przejściowe wstępnego filtru dolnoprzepustowego poszerza się, a zwęża się pasmo niespecyfikowane filtrów ułamkowo-opóźniających. Zatem w takim przypadku rośnie wymagana długość odpowiedzi impulsowej filtrów ułamkowo-opóźniających N_{FSD} a maleje wymagana długość odpowiedzi wstępnego filtru dolnoprzepustowego N_{LPF} . Odpowiedni dobór tego parametru pozwala uzyskać np. rozwiązanie o minimalnych wymaganiach pamięciowych lub kosztach numerycznych. W celu przyspieszenia procesu doboru tego parametru można posłużyć się wzorami szacunkowymi na długość odpowiedzi impulsowych [3,14].

Rozpatrzmy teraz przykładowy projekt algorytmu konwersji z 48 000 próbek/s na 44 100 próbek/s. Założymy odchylenie charakterystyki amplitudowej poniżej częstotliwości 19 kHz nie większe niż 0.1dB, a tłumienie składowych mogących spowodować zniekształcenia aliasingowe nie mniejsze niż 100dB. Przyjmiemy parametr $\alpha = 0.567$ minimalizujący koszty numeryczne przy założeniu, że wszystkie filtry ułamkowo-opóźniające zaprojektujemy zawczasu i będziemy ich współczynniki pobierać z tablicy LUT.

Okazuje się, że aby spełnić powyższe wymagania, potrzebny jest wstępny filtr dolnoprzepustowy o długości $N_{LPF} = 61$ i 147 filtrów ułamkowo opóźniających o długości $N_{FSD} = 68$. Jak widać filtry te są znacznie krótsze od filtru w rozwiązaniu klasycznym i zaprojektowanie ich nie stwarza żadnych trudności technicznych. Razem mamy aż 10 057 wszystkich współczynników filtrów, które należy przechować w tablicy LUT. Jednak jest to wielkość nieznacznie tylko większa od szacowanej liczby 9 413 współczynników filtru interpolacyjno-decymacyjnego dla klasycznego rozwiązania.



Rys.5. Charakterystyki zbiorcze jednostopniowego algorytmu zmiany szybkości próbkowania.

Założenia projektowe: $f_{p1}=48\text{kHz}$, $f_{p2}=44.1\text{kHz}$, $f_{max} = 19\text{kHz}$, $A_p = 0.1\text{dB}$, $A_s = 100\text{dB}$.

Charakterystyki zbiorcze algorytmu zmiany szybkości próbkowania zaprojektowanego dla przyjętych założeń przedstawiono na rys. 5. Charakterystykę zbiorczą algorytmu

przepróbkowania wykreślono tutaj grubą czarną linią. Wykres umieszczony w „tle” wyskalowany od 0 do 3.5 MHz umieszczono tutaj by pokazać, jak bardzo wąskopasmowy musi być filtr zbiorczy i jak ostre są postawione wymagania. By można było określić w jakim stopniu spełnione są postawione wymagania, na rysunku umieszczono dodatkowo wykres charakterystyk wyskalowany od 0 do 88.2 kHz (zaznaczono ten fragment na rysunku w „tle” szarym prostokątem) oraz powiększony fragmentu charakterystyk z przedziału $(0, f_{\max})$. Szarą linią wykreślono charakterystykę zbiorczą algorytmu przepróbkowania dla przypadku pominięcia wstępnego filtra dolnoprzepustowego i dodatkowo cienką czarną linią naniesiono charakterystykę wstępnego filtra dolnoprzepustowego (wraz z replikami wokół wielokrotności częstotliwości f_{p1}). Na wykresach wyraźnie widać, że wszystkie postawione wymagania są spełnione.

4. ROZWIĄZANIE DWUSTOPNIOWE

Zaprezentowane w poprzednim punkcie rozwiązanie spełnia postawione wymagania i co ważniejsze, wykonanie takiego projektu jest praktycznie realizowalne. Jednak w niektórych zastosowaniach duże wymagania pamięciowe mogą być poważnym mankamentem. Okazuje się jednak, że problem ten można rozwiązać stosując dwustopniową realizację procesu zmiany szybkości próbkowania.

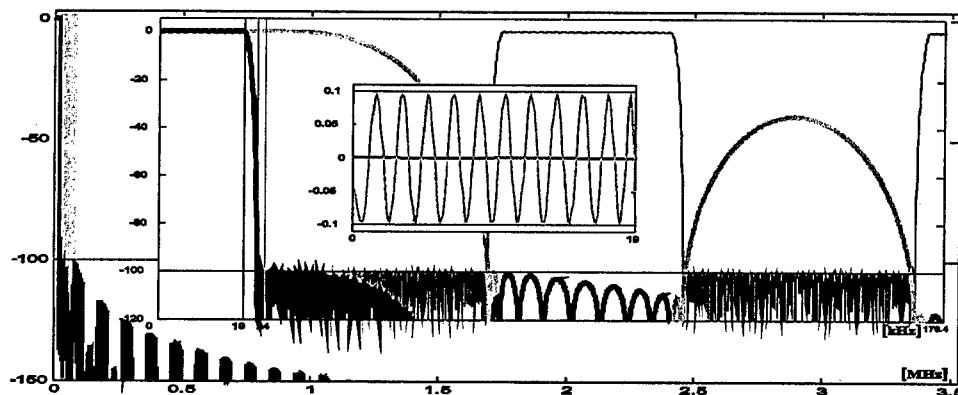
Zwróćmy tutaj uwagę na to, że wymagania pamięciowe wynikają głównie z konieczności przechowywania współczynników filtrów ułamkowo-opóźniających. Okazuje się jednak, że realizując przepróbkowanie dwustopniowo: najpierw zwiększając szybkość próbkowania, np. dwukrotnie, a dopiero potem przepróbkując taki sygnał do docelowej szybkości próbkowania, można uzyskać znaczne zmniejszenie wymagań pamięciowych algorytmu.

W rozwiązaniu tym zadania wstępnego filtra dolnoprzepustowego może realizować filtr interpolacyjny pierwszego stopnia dwustopniowego algorytmu przepróbkowania. Przy wstępnej dwukrotnej interpolacji i przy założeniach jak w p. 3, długość odpowiedzi impulsowej filtra interpolacyjnego dla pierwszego stopnia algorytmu wynosi $N_{LPF} = 118$. Współczynników tych jest tylko dwa razy więcej niż dla wstępnego filtra w rozwiązaniu jednostopniowym.

Należy jeszcze zaprojektować drugi stopień algorytmu przepróbkowania, realizujący konwersję z szybkości $2f_{p1}$ na szybkość f_{p2} . Parametry L i M drugiego stopnia wynoszą, odpowiednio, 147 i 320, co oznacza, że wymagana liczba filtrów ułamkowo-opóźniających nie uległa zmianie i wynosi 147. Filtry ułamkowo-opóźniające będą pracowały teraz z dwa razy większą szybkością, a przetwarzany sygnał nie jest już sygnałem nienadpróbkowanym. W wyniku wstępnej interpolacji uzyskujemy bowiem sygnał dwukrotnie nadpróbkowany. W efekcie, przy tej samej liczbie filtrów ułamkowo-opóźniających, długość ich odpowiedzi impulsowych jest tym razem prawie 7-krotnie mniejsza i wynosi $N_{FSD} = 10$. Stąd też tablica LUT musi tym razem pomieścić tylko 1588 współczynników. Jednocześnie koszty numeryczne implementacji tej wersji algorytmu rosną tylko nieznacznie, przy założeniu, że pierwszy stopień interpolacji zrealizujemy w oparciu o strukturę polifazową. Średnia wymagana liczba mnożeń i dodawań w przeliczeniu na jedną próbkę wyjściową wynosi dla tego algorytmu 148.4, podczas gdy dla rozwiązania jednostopniowego wielkość ta wynosi 134.4.

Charakterystyki zbiorcze algorytmu dwustopniowego przedstawiono na rys. 6. Charakterystykę zbiorczą algorytmu przepróbkowania wykreślono tutaj grubą czarną linią. Układ rysunku jest taki sam jak rys. 5, z tą tylko różnicą, że aby pokazać artefakty filtra

zbiorczego drugiego stopnia algorytmu, powiększony fragment charakterystyk wyskalowano od 0 do 176.4 kHz. Cienką linią naniesiono tutaj charakterystykę filtra interpolacyjnego pierwszego stopnia algorytmu (wraz z replikami wokół wielokrotności częstotliwości $4f_{p1}$). Na wykresach możemy zobaczyć, że wszystkie postawione wymagania są spełnione.



Rys.6. Charakterystyki zbiorcze dwustopniowego algorytmu zmiany szybkości próbkowania.

Założenia projektowe: $f_{p1}=44.1\text{kHz}$, $f_{p2}=48\text{kHz}$, $f_{max} = 18\text{kHz}$, $A_p = 0.2\text{dB}$, $A_s = 92\text{dB}$

5. PODSUMOWANIE

W pracy omówiono klasyczne podejście do implementacji zmiany szybkości próbkowania sygnałów dyskretnych. W szczególności pokazano problemy konwersji pomiędzy telekomunikacyjnymi i multimedialnymi szybkościami próbkowania. Wysoka niewspółmierność tych szybkości próbkowania sprawia, że w przypadku klasycznego algorytmu przepróbkowania konieczne jest zaprojektowanie filtra dolnoprzepustowego o bardzo wąskim pasmie przepustowym i szczególnie długiej odpowiedzi impulsowej. W efekcie w tego typu zastosowaniach należy znaleźć inne rozwiązanie. Jako alternatywne rozwiązanie zaproponowano tutaj algorytm pracujący w oparciu o filtry ułamkowo-opóźniające. W rozwiązaniu tym jeden filtr interpolacyjno-decymacyjny o bardzo długiej, rzędu kilku tysięcy próbek, odpowiedzi impulsowej zastępuje się wieloma filtrami ułamkowo-opóźniającymi o odpowiedziach impulsowych nieprzekraczających kilkudziesięciu próbek. Jednak, aby skutecznie zaprojektować wysokiej jakości algorytm przepróbkowania konieczna jest możliwość przeniesienia wymagań dotyczących jakości algorytmu konwersji, na specyfikacje filtrów ułamkowo-opóźniających. Dlatego też w pracy proponujemy zastosowanie filtrów ułamkowo-opóźniających optymalnych w sensie Czebyszewa, dla których możliwe jest określenie reguł przeniesienia tych wymagań. By zaprezentować skuteczność prezentowanej koncepcji, przedstawiono dwa przykłady projektów algorytmów konwersji szybkości próbkowania z telekomunikacyjnej (48 000 próbek/s) na multimedialną (44 100 próbek/s). Drugi z przykładów prezentuje możliwość znacznego zmniejszenia wymagań pamięciowych implementacji algorytmu konwersji poprzez zastosowanie wstępnej interpolacji. Ponadto zastosowanie rozwiązania dwustopniowego wiąże się tylko z nieznacznym zwiększeniem złożoności numerycznej algorytmu.

BIBLIOGRAFIA

- [1] Lim J.S., Oppenheim A.V. (Eds.): *Advanced Topics in Signal Processing*, Prentice Hall, 1988.
- [2] Mitra S.: *Digital Signal Processing: A computer-based approach*, McGraw-Hill, 1998.
- [3] Zelniker G., Taylor F.J.: *Advanced Digital Signal Processing. Theory and Applications*, Marcel Dekker, New York, 1994.
- [4] Marciniak T., Dąbrowski A.: *Multi-Channel Digital Conversion of Audio Sampling Rates*, Online Symposium for Electronics Engineers, TechOnLine Bedford, Massachusetts (USA), 01 December 2001, http://www.techonline.com/community/ed_resource/feature_article/14778.
- [5] Laakso T.I., Valimäki V., Karjalainen M., and Laine U.K.: *Splitting the unit delay*. IEEE Signal Processing Magazine, vol.13 (No.1), pp. 30-60, January 1996.
- [6] Valimäki V., Laakso T.I.: *Principles of Fractional Delay Filters*, ICASSP 2000 Istanbul, Turkey, June 5-9.
- [7] Tarczyński A., Koziński W., Cain G.D.: *Sampling rate conversion using fractional-sample delay*, ICASSP'94, 1994, vol. 3, pp. III-285 – III-288.
- [8] Murphy N.P., Tarczyński A., Laakso T.I.: *Sampling-rate conversion using a wideband tuneable fractional delay element*, NORSIG'96, 1996, pp. 423-426.
- [9] Hermanowicz E., Rojewski M., Blok M.: *A sample rate converter based on a fractional delay filter bank*, ICSPAT 2000, Dallas, Tx, USA, October 16-19.
- [10] Blok M.: *Algorytm próbkowania sygnałów pomiędzy standardami CD i DAT*, KST'2001, Bydgoszcz, tom B, str. 258-267.
- [11] Blok M.: *Przepróbkowanie sygnałów cyfrowych realizowane za pośrednictwem filtrów FSD optymalnych w sensie kryterium LS*, KST'2002, Bydgoszcz, tom B, str. 43-52.
- [12] Blok M.: *Collective Filter Evaluation of an FSD Filter-Based Resampling Algorithm*, Online Symposium for Electronics Engineers, TechOnLine Bedford, Massachusetts (USA), 15 January 2002, http://www.techonline.com/community/ed_resource/feature_article/14917.
- [13] Blok M.: *Projektowanie opóźniających filtrów cyfrowych FIR metodą iteracji czasowo-częstotliwościowej*, Rozprawa doktorska, Politechnika Gdańska, WETI, Gdańsk 2003.
- [14] Blok M.: *Szacowanie długości ułamkowo-opóźniających filtrów typu FIR*, KST'2000, Bydgoszcz, tom A, str. 325-332.

DIGITAL SIGNALS RESAMPLING BETWEEN TELECOMMUNICATION AND MULTIMEDIA SAMPLING RATES

Summary

In this paper the problem of sampling rate conversion in application to interfacing systems working with telecommunication and multimedia sampling rates is discussed. The classic approach to the resampling is presented and its faults concerning given application are demonstrated. As an alternative, the approach based on fractional sample delay filters optimal in the Chebyshev sense is proposed. Its applicability to high quality sampling rate conversion is presented on the basis of two design examples. The second example presents a novel two-stage implementation that results in a significant memory requirements reduction.

Marek Blok, Mirosław Rojewski, Adam Sobociński

Katedry Systemów Informacyjnych, Politechnika Gdańska

NOWY ESTYMATOR TONU KRTANIOWEGO

Streszczenie

W pracy przedstawiono nową koncepcję potokowego algorytmu estymacji (estymatora) wartości chwilowej częstotliwości podstawowej tzw. tonu krtaniowego sygnału mowy. Estymator ten wykorzystuje bank zespolonych filtrów wąskopasmowych oraz dwupunktowy estymator pulsacji chwilowej bez rozwijania fazy. Przebiegi zespolone otrzymywane na wyjściach tych filtrów, których pasma zostały dopasowane do właściwości sygnału mowy, są poddawane demodulacji częstotliwości i amplitudy przy użyciu estymatora pulsacji chwilowej i obwiedni, a następnie na podstawie tych zdemodulowanych przebiegów w części decyzyjnej algorytmu dokonywany jest wybór tej pulsacji, która odpowiada szukanej wartości częstotliwości chwilowej tonu krtaniowego. W pracy zamieszczono wyniki eksperymentów przeprowadzonych na sygnale mowy przy użyciu komputerowej implementacji proponowanego rozwiązania.

1. WSTĘP

Przedmiot naszych rozważań, tzw. ton krtaniowy, jest fizycznym (akustycznym) wynikiem drgań fałd (strun) głosowych człowieka podczas wypowiadania dźwięków mowy zwanych samogłoskami. Ton krtaniowy jest (prawie) okresowym ciągiem impulsów krtaniowych, które stanowią tzw. pobudzenie krtaniowe traktu głosowego podczas artykulacji samogłosek. Najważniejszym parametrem tonu krtaniowego jest jego wysokość (ang. *pitch*) mierzona częstotliwością powtarzania impulsów krtaniowych (ang. *pitch frequency*) albo ich okresem (ang. *pitch period*). Pomiar tego parametru w oparciu o dostępny sygnał mowy nazywamy estymacją tonu krtaniowego (ang. *pitch estimation/determination/extraction/tracking*). Wynikiem estymacji tonu krtaniowego jest jego estymata w postaci przebiegu częstotliwości chwilowej w hercach [Hz] (w teoretycznych rozważaniach – unormowana pulsacja chwilowa w radianach na próbkę [rad/Sa]) albo w postaci przebiegu okresu chwilowego w milisekundach [ms]. Wykres estymaty tonu krtaniowego odpowiadający intonacji wypowiadanej samogłoski nazywa się profilem tonu krtaniowego (ang. *pitch profile/contour/pattern*). Do wypreparowania profili tonu krtaniowego z potokowo (ang. *on-line*) otrzymanej estymaty służy detektor mowy dźwięcznej (ang. *voiced detector*), którym w tej pracy nie zajmujemy się [1].

Rozwiązanie problemu estymacji tonu krtaniowego sygnału mowy odgrywa bardzo istotną rolę w obszarach techniki związanych z analizą tego sygnału, takich jak kodowanie,

synteza, rozpoznawanie i rekonstrukcja sygnału mowy [1]. Trudności związane z estymacją tonu krtaniowego sygnału mowy wynikają ze zmienności i nieregularności tego sygnału. Struktura harmoniczna sygnału mowy ulega ciągłym zmianom w zależności od artykulacji różnych dźwięków, a występowanie głosek bezdźwięcznych, które nie tworzą struktur „okresowych”, dodatkowo komplikuje zadanie. W związku z tym, że impulsy krtaniowe są wydobywane z różną siłą, nawet wtedy, gdy występują bezpośrednio po sobie, sygnał jest zmodulowany amplitudowo (ang. *shimmer*). Odstępy między kolejnymi impulsami krtanowymi w ramach jednej głoski dźwięcznej również mogą ulegać zmianie (ang. *jitter*).

Wśród stosowanych metod estymacji tonu krtaniowego można wyodrębnić kilka grup. Pierwszą grupę stanowią te, w których stosuje się analizę częstotliwościową z użyciem FFT. W takim przypadku problem sprowadza się do wyboru transformaty FFT o odpowiedniej długości/rozdzielczości, a następnie, wykrycia, który z prążków obliczonej transformaty odpowiada tonowi krtaniowemu [2], ewentualnie gdy sygnał foniczny jest odfiltrowany od dołu (np. bas przez telefon), okres tonu krtaniowego określamy z odstępów prążków harmonicznych. Inną grupę stanowią algorytmy działające w oparciu o bank filtrów służących wydobywaniu podpasem częstotliwościowych, w których następuje właściwa estymacja częstotliwości chwilowej [3]. Często stosowane są także metody korelacyjne, które polegają na określeniu stopnia podobieństwa między fragmentami sygnału w przesuniętych względem siebie oknach czasowych [4,5,6].

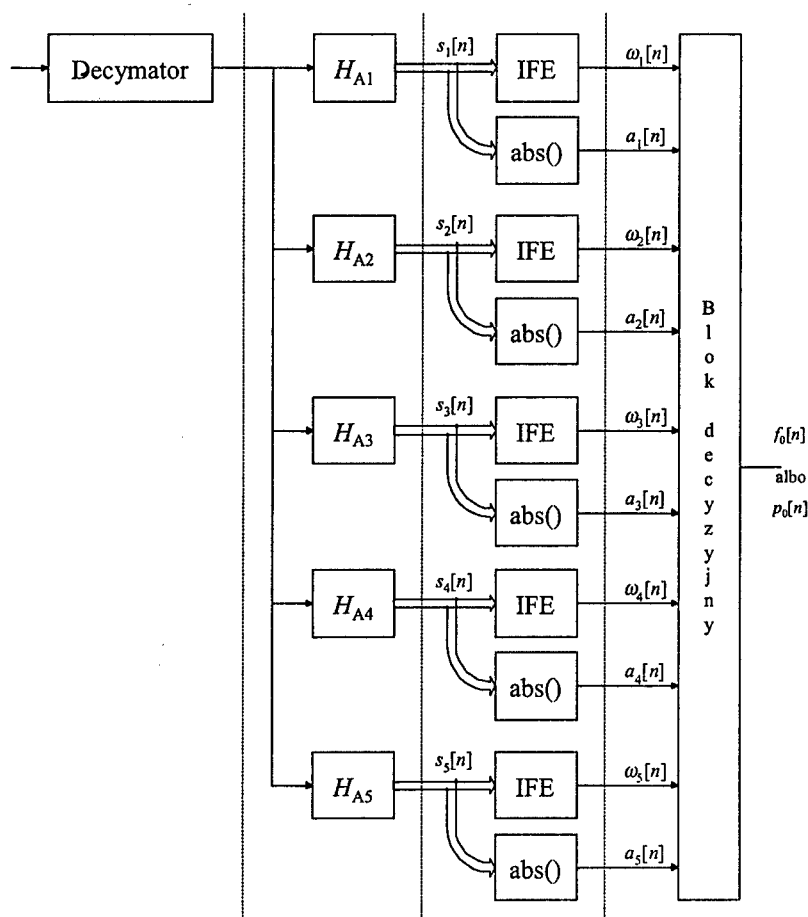
Prezentowany w tej pracy algorytm należy do drugiej grupy. Zastosowany w nim został bank pięciu filtrów zespolonych o parametrach dopasowanych do sygnału mowy zawierającego w swoim widmie prążek podstawowy, tzw. f_0 . Bank ten „pokrywa” zakres częstotliwości tonu krtaniowego głosu męskiego i kobiecego. W każdym z uzyskanych podpasem częstotliwościowych dokonywana jest estymacja częstotliwości chwilowej oraz amplitudy chwilowej. Do estymacji częstotliwości wykorzystywany jest bardzo krótki, dwupunktowy estymator pulsacji chwilowej bez rozwijania fazy [7]. Na podstawie amplitudy chwilowej podejmowana jest decyzja, która spośród pięciu otrzymanych wartości częstotliwości chwilowej jest szukaną częstotliwością chwilową tonu krtaniowego.

Dalej, w p.2, opisujemy dokładniej proponowany estymator tonu krtaniowego, następnie, w p.3, przytaczamy przykład zastosowania nowego algorytmu i porównujemy jego estymatę z estymatą otrzymaną metodą autokorelacyjną dla tej samej frazy. Na koniec, w p.4, podsumowujemy pracę.

2. KONCEPCJA NOWEGO ESTYMATORA TONU KRTANIOWEGO

Działanie proponowanego estymatora tonu krtaniowego można podzielić na kolejne etapy (rys. 1). W pierwszym etapie następuje decymacja przetwarzanego sygnału, której celem jest obniżenie szybkości próbkowania, a przez to znaczne zmniejszenie kosztów numerycznych algorytmu. Możliwość wykonania tej operacji wynika z tego, że maksymalna wartość szukaney częstotliwości chwilowej dla sygnału mowy nie powinna raczej przekroczyć 300-400 Hz, a więc zupełnie wystarczająca jest szybkość próbkowania $F_s = 1000$ Sa/s (próbek na sekundę), podczas gdy standardowe szybkości próbkowania wynoszą od 8000 do 48000 Sa/s. Drugi etap polega na przefiltrowaniu zdecydowanego sygnału przy użyciu banku filtrów, które zostały specjalnie zaprojektowane dla sygnału mowy, tzn. są to filtry półoktawowe (każdy następny, tj. o wyższej częstotliwości środkowej, filtr ma pasmo $\sqrt{2}$ razy szersze niż poprzedni), obejmujące zakres częstotliwości odpowiadających wartościom częstotliwości podstawowej tonu krtaniowego, od naj-

niższego głosu męskiego (bas) do najwyższego głosu kobiecego, uzyskując „pokrycie częstotliwości” od 50 do 400 Hz. Specyfikację banku tych filtrów prezentujemy dalej na rys. 3. W wyniku przeprowadzenia tej filtracji uzyskuje się pięć wąskopasmowych sygnałów zespolonych. W trzecim etapie przeprowadzana jest estymacja częstotliwości chwilowej i amplitudy chwilowej otrzymanych sygnałów zespolonych. Na podstawie wartości amplitud chwilowych w poszczególnych podpasmach podejmowana jest decyzja, która z otrzymanych częstotliwości chwilowych jest szukaną częstotliwością podstawową tonu krtaniowego.



Rys.1. Ogólna koncepcja proponowanego algorytmu estymacji częstotliwości chwilowej albo okresu chwilowego tonu krtaniowego.

Z punktu widzenia projektowania algorytmu CPS proces estymacji częstotliwości chwilowej albo okresu chwilowego tonu krtaniowego w algorytmie zaproponowanym w tej pracy jest kaskadą trzech etapów (rys. 1). Pierwszy etap to ograniczenie pasma sygnału mowy do interesującego nas zakresu połączone z decymacją, co pozwala znacznie ograniczyć koszty numeryczne przetwarzania. Drugi etap to rozfiltrowanie sygnału za pomocą banku zespolonych filtrów wąskopasmowych na pięć składowych. W trzecim etapie dla

każdego z sygnałów otrzymanych na wyjściu banku filtrów wykonywana jest estymacja amplitudy chwilowej i częstotliwości chwilowej i na podstawie tych przebiegów jest podejmowana ostateczna decyzja.

2.1 Etap decymacji

Pierwszym etapem na drodze do estymacji częstotliwości tonu podstawowego jest wstępne obniżenie szybkości próbkowania do 1000 Sa/s (rys. 1). Z jednej strony pozwala to na niezależne od szybkości próbkowania przetwarzanego sygnału projektowanie podstawowej części algorytmu występującej po decymatorze. Z drugiej zaś, przede wszystkim skutkuje wyeliminowaniem z przetwarzanego sygnału składowych nas nie interesujących (pasmo sygnału zostaje ograniczone do 500Hz), a to pozwala uprościć zagadnienie projektowania i obniżyć koszty numeryczne implementacji banku zespolonych filtrów wąskopasmowych.

Rozpatrzmy teraz zagadnienie projektowania tegoż decymatora. W przypadku konwersji z szybkości próbkowania 8000 Sa/s do szybkości 1000 Sa/s. Wymagana jest tutaj 8-krotna decymacja, która jest stosunkowo łatwa i tania w implementacji. Jednak, gdy chcemy sygnał zddecymować do tej samej szybkości próbkowania z szybkości 44100 Sa/s, napotykamy na bardzo trudne zadanie przepróbkowania sygnału w stosunku 441:10. Dlatego też bliżej omówimy tutaj właśnie ten trudniejszy przypadek.

Proponujemy tutaj zastosować prostszą, 44-krotną decymację, oraz dodatkowo w celu ułatwienia projektowania potrzebnych filtrów decymacyjnych i zmniejszenia kosztów implementacji numerycznej rozpatrzmy implementację dwustopniową: najpierw wykonujemy decymację 11-krotną, a następnie 4-krotną. W efekcie uzyskamy szybkość próbkowania $1002 + 3/11$ Sa/s, nieznacznie tylko różniącą się od wartości pożądanej. Zauważmy, że różnica ta jest na tyle nieznaczna, że w praktyce niezależnie od tego, czy wejściowa szybkość próbkowania wynosi 8000 czy 44100 kSa/s występujący po decymatorze bank zespolonych filtrów wąskopasmowych nie wymaga odrębnego projektowania. Różnice w szybkości próbkowania należy jedynie uwzględnić na wyjściu całego estymatora pulsacji, przy przeliczaniu pulsacji chwilowej tonu podstawowego w rad/Sa na częstotliwość w Hz.

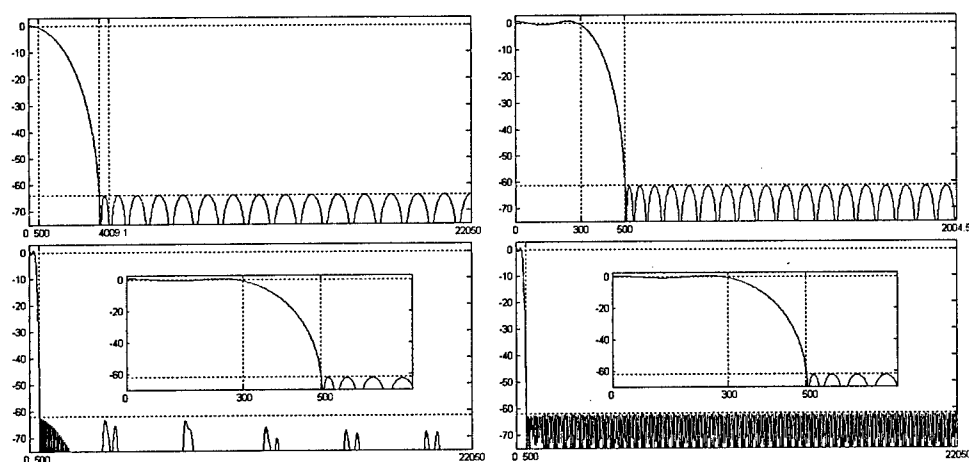
Przy projektowaniu decymatora założyliśmy, że powinien on bez większych zniekształceń przenosić pasmo do 300Hz oraz, że składowe mogące spowodować zniekształcenia aliasowe powinny być wytłumione na poziomie co najmniej 60dB. Dodatkowo, by dopasować filtr decymacyjny do wydajnej numerycznej implementacji polifazowej założyliśmy, że długość odpowiedzi impulsowej filtru decymacyjnego powinna być całkowitą wielokrotnością krotności decymacji. Przy takich założeniach filtry zaprojektowaliśmy za pomocą algorytmu Remeza generującego filtry optymalne w sensie kryterium MINIMAX [8].

Charakterystyki filtru spełniającego powyższe wymagania dla jednostopniowej implementacji decymatora przedstawia rys. 2 (na dole po prawej). Długość odpowiedzi impulsowej tego filtru wynosi aż $11 \cdot 44 = 484$ Sa. W przypadku implementacji polifazowej decymator ten wymaga 484-ech mnożeń i dodawań na jedną próbkę wyjściową.

W rozwiązaniu dwustopniowym pierwszy stopień realizuje decymację 11-krotną, a drugi 4-krotną. W pierwszym stopniu, przy założeniu, że nie dopuszczamy zniekształceń aliasowych w pasmie do 500Hz, pasmo przejściowe filtru można ulokować od 500Hz do 3509Hz ($44100/11 - 500$). Długość odpowiedzi impulsowej filtru spełniającego nasze wymagania to zaledwie $3 \cdot 11 = 33$ Sa. W przypadku filtru decymacyjnego dla drugiego stopnia decymacji pasmo przejściowe, tak samo jak dla wersji jednostopniowej, należy

ulożować pomiędzy 300Hz a 500Hz, jednak tym razem wejściowa szybkość próbkowania jest znacznie mniejsza. W efekcie długość odpowiedzi impulsowej zaprojektowanego filtra wynosi $11 \cdot 4 = 44$ Sa, a cały dwustopniowy decymator wymaga 176-ciu mnożeń i dodawań na próbkę wyjściową. Charakterystyki filtrów obu filtrów decymacyjnych przedstawiono na rys. 2 (u góry), gdzie umieszczono również (na dole po lewej) charakterystykę zbiorczą (odpowiednik charakterystyki filtra decymacyjnego w algorytmie jednostopniowym) tej implementacji decymatora.

Dodatkową zaletą rozwiązania dwustopniowego jest to, że jedynie wymieniając albo pomijając drugi stopień decymatora można łatwo przystosować proponowane rozwiązanie do analizy sygnałów spróbkowanych z innymi szybkościami np. 22050 czy 11025 Sa/s.



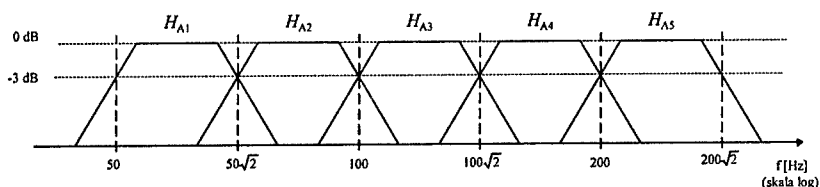
Rys.2. Charakterystyki amplitudowe filtrów decymacyjnych dla wejściowej szybkości próbkowania 44100 Sa/s; dla decymatora dwustopniowego z pierwszego stopnia (u góry po lewej) oraz drugiego stopnia (u góry po prawej). Na dole, po lewej, charakterystyka zbiorcza decymatora dwustopniowego a po prawej, charakterystyka filtra decymacyjnego dla jednostopniowej implementacji decymatora. Oś pionowa w dB, oś pozioma w Hz.

2.2 Bank zespolonych filtrów wąskopasmowych

Kluczowym elementem zaproponowanego tutaj algorytmu jest bank zespolonych filtrów wąskopasmowych. Ich zadaniem jest zapewnienie dobrych warunków pracy dla występujących na ich wyjściu estymatorów pulsacji chwilowej. By estymator pulsacji chwilowej omówiony w p.2.3 mógł spełnić swoje zadanie, na jego wejście musi być podany zespolony ciąg hilbertowski („analityczny”) zawierający tylko jedną składową częstotliwościową analizowanego przebiegu okresowego.

By uzyskać zamierzony efekt proponujemy zastosować bank zespolonych filtrów półoktawowych o pasmach wyspecyfikowanych na rys. 3 pokrywających z nadmiarem zakres typowych częstotliwości podstawowych drgania krtaniowego. Ponadto, ponieważ ciągi na wyjściu filtrów mają być hilbertowskie, zakładamy zerową charakterystykę amplitudową dla ujemnych częstotliwości. Istotą specyfikacji z rys. 3 jest to, że jeżeli składowa podstawowa tonu krtaniowego znajduje się w pasmie filtra H_{Ai} ($i=1,2,3$ a indeks A do podkreślenia „analityczności” sygnałów na wyjściu tych filtrów) to jej harmoniczna pojawia się poza pasmem tego filtra, w pasmie filtra H_{Ai+2} . Przy projektowaniu tych filtrów

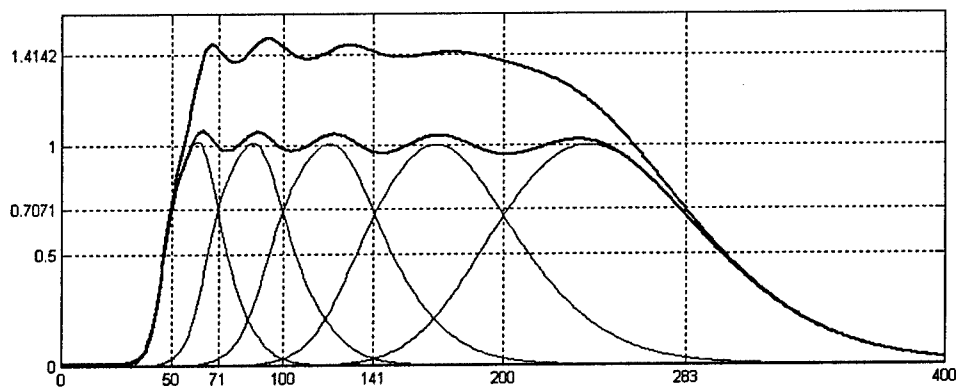
ważna jest selektywność zapewniająca odpowiednią czystość widmową sygnału na wyjściu filtru oraz komplementarność amplitudowa zapewniająca równoważne traktowanie sygnałów pojawiających się w pasmach różnych filtrów. Sam kształt charakterystyk w pasmie przepustowym jest mniej istotny. Do tego celu najlepiej nadają się filtry typu IIR o stałej dobroci (ang. *constant-Q*).



Rys.3. Specyfikacja banku zespolonych filtrów wąskopasmowych w skali logarytmicznej

Przy projektowaniu banku zespolonych filtrów o stałym Q dąży się do symetrii charakterystyk częstotliwościowych na skali logarytmicznej wokół częstotliwości środkowej liczonej jako średnia geometryczna z częstotliwości brzegowych. Jednym słowem zbocze niskoczęstotliwościowe jest tu zawsze bardziej strome od zbocza wysokoczęstotliwościowego. Za najlepszą prototypową charakterystykę amplitudową pojedynczego filtru z *constant-Q* banku uważa się powszechnie krzywą gaussowską na skali „log-lin”, której odpowiada charakterystyka „minus kwadratowa” na skali „log-log” (zwanej też skalą „log-dB”).

Dla naszych potrzeb zaprojektowaliśmy bank filtrów IIR szóstego rzędu aproksymujących charakterystyki spełniających określone powyżej wymagania. Charakterystyki amplitudowe poszczególnych filtrów z banku oraz charakterystyki komplementarności amplitudowej i mocy są wykreślone na rys. 4.



Rys.4. Charakterystyki amplitudowe zespolonych filtrów wąskopasmowych (cienka linia) oraz charakterystyki komplementarności amplitudy (linia ustalająca się na poziomie $\sqrt{2}$) i mocy (linia ustalająca się na poziomie 1).

2.3 Estymacja częstotliwości chwilowej i amplitudy chwilowej

Do estymacji pulsacji chwilowej użyto prostego algorytmu (estymatora pulsacji chwilowej, na rys. 1 oznaczony jako IFE – *instantaneous frequency estimator*), wg którego pulsację chwilową sygnału hilbertowskiego („analitycznego”) $s_i[n]$ na wyjściu i -tego filtru wąskopasmowego H_{Ai} , $i=1,2,\dots,5$ obliczamy ze wzoru definicyjnego

$$\omega_i[n] = \text{Arg}(s_i[n]s_i^*[n-1]) \quad (2.1)$$

We wzorze tym n oznacza bieżący numer próbki, $*$ – sprzężenie zespolone, zaś $\text{Arg}()$ jest „arcustangensem czteroćwiartkowym” aproksymowanym w MATLABie funkcją $\text{angle}()$. Jak wiadomo [7] estymator ten oblicza wartość główną pulsacji chwilowej $\omega_i[n] \in [-\pi, \pi)$ rad/Sa i przy uniwersalnym jego zastosowaniu należałoby ciąg $\omega_i[n]$ poddać rozwijaniu poza ten przedział. Jednakże ze względu na wąskopasmowość i w dominującej większości jednotonowość sygnałów $s_i[n]$, prawdopodobieństwo wykroczenia $\omega_i[n]$ poza okres ich wartości głównej należy uznać za znikomo małe nawet dla szumowego pobudzenia tych filtrów. Z tego powodu zdecydowaliśmy się na tę uproszczoną postać estymatora IFE.

Z uwagi na to, że IFE nie wprowadza opóźnień, możemy równie prosto, synchronicznie z obliczaniem $\omega_i[n]$, obliczyć amplitudę chwilową jako

$$a_i[n] = |s_i[n]| \quad (2.2)$$

Wykonuje to funkcja $\text{abs}()$ na rys. 1.

Uzyskana w ten sposób para przebiegów rzeczywistych: $a_i[n]$ i $\omega_i[n]$ stanowi reprezentację AM·FM [1] i -tego zespolonego sygnału wąskopasmowego $s_i[n]$. Wszystkie pięć reprezentacji AM·FM jest przekazywane dalej do bloku decyzyjnego.

2.4 Blok decyzyjny

Ostatnim, bardzo ważnym, elementem zaproponowanego algorytmu jest blok decyzyjny, którego zadaniem jest określenie, w paśmie którego filtru znajduje się w danej chwili składowa podstawowa tonu krtaniowego. Inaczej mówiąc, w bloku tym podejmowana jest decyzja, którą z pięciu pulsacji chwilowych $\omega_i[n]$ uznać za pulsację tonu krtaniowego w chwili n , oznaczoną dalej jako $\omega_0[n]$. Decyzja ta podejmowana jest na podstawie porównania wielkości amplitud chwilowych $a_i[n]$ towarzyszących poszczególnym pulsacjom w reprezentacji AM·FM obserwowanych na wyjściu każdego z filtrów. Porównanie to może być złożonym procesem, my jednak rozpatrzmy tutaj najprostsze podejście, które okazuje się często skuteczne (co można zaobserwować, na „mapie” pulsacji chwilowych na rys. 5), w którym za właściwy filtr uważamy filtr na wyjściu którego w danej chwili obserwujemy największą amplitudę chwilową. A więc

$$\omega_0[n] = \omega_{i_{\max}}[n] \mid a_{i_{\max}}[n] > a_i[n] \forall i \neq i_{\max} \quad (2.3)$$

W powyższym zapisie i_{\max} oznacza indeks i , dla którego amplituda $a_i[n]$ w chwili n jest maksymalna, a kreska pionowa oznacza warunek. Zakładamy tutaj, że filtr, na którego wyjściu obserwujemy przebieg zespolony $s_i[n]$ o największej wartości amplitudy chwilowej $a_i[n]$, jest tym, w którego paśmie znajduje się składowa podstawowa tonu krtaniowego o pulsacji $\omega_0[n]$ obliczonej wg (2.1)-(2.3).

Otrzymańmy w ten sposób pulsację chwilową tonu krtaniowego przeliczany na jego częstotliwość chwilową $f_0[n]$ albo na jego okres chwilowy $p_0[n]$ wg wzorów:

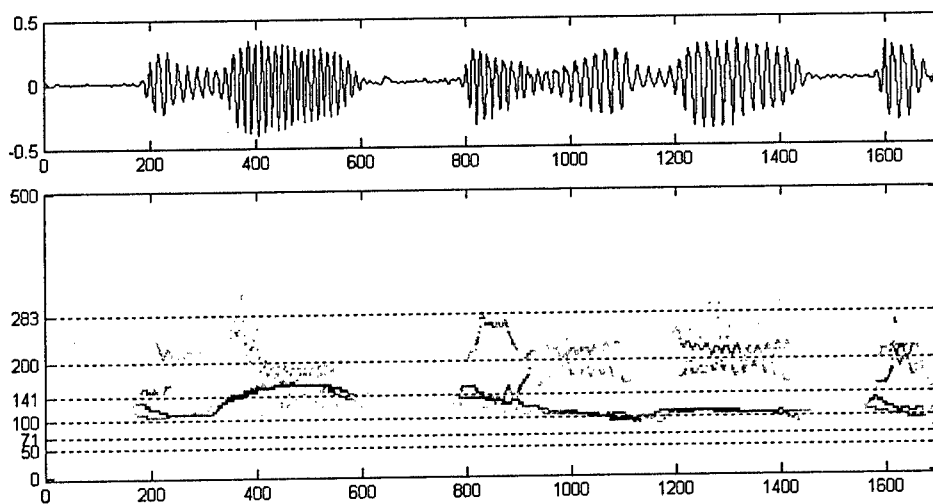
$$f_0[n] = F_s \omega_0[n] / (2\pi), \quad p_0[n] = 1 / f_0[n] = 2\pi / (F_s \omega_0[n]) \quad (2.4)$$

gdzie F_s jest szybkością próbkowania analizowanego sygnału mowy po decymacji.

3. PRZYKŁADOWE POMIARY

By przedstawić skuteczność działania zaproponowanej metody na rys. 5 i 6 pokazano wyniki pomiarów dla frazy „*studenckiej agencji*” wypowiedzianej męskim głosem i spróbkowanej z szybkością $F_s = 44100$ Sa/s. Dla porównania, na rysunkach tych umieszczono również przebiegi uzyskane za pomocą uważanej za dokładną ale za to znacznie bardziej złożonej numerycznie, metody bazującej na obliczaniu autokorelacji analizowanych fragmentów [4,5,6].

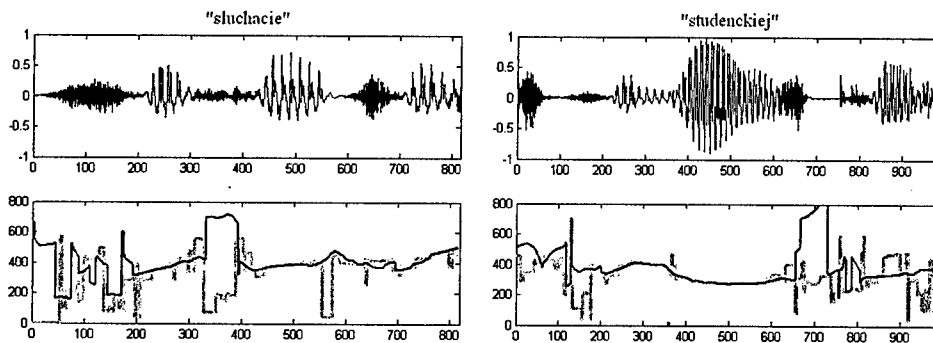
Na rys. 5 zilustrowano pośrednie przebiegi uzyskiwane w trakcie pracy zaproponowanego algorytmu dla frazy „*studenckiej agencji*”. Ilustrują one działanie poszczególnych etapów przetwarzania. Na górnym wykresie wykreślono sygnał z wyjścia decymatora, który podawany jest na wejście wąskopasmowego banku filtrów zespolonych. Jeżeli przebieg ten porównać z odpowiednimi fragmentami sygnału mowy z rys. 6 można zauważyć, że przebieg ten jest „znacznie łagodniejszy” (jego pasmo zostało istotnie ograniczone), a występujące we frazie głoski bezdźwięczne (np. „st” na rys. 5 fragment do 200ms i „c” fragment od 1450ms do 1580ms) uległy prawie kompletnemu wytłumieniu.



Rys.5. U góry zdecymowany do 1000 Sa/s sygnał mowy z wejścia banku zespolonych filtrów wąskopasmowych (faza „*studenckiej agencji*”) oraz na dole „mapa” częstotliwości chwilowych przebiegów otrzymanych na wyjściu banku filtrów. Oś pozioma obu rysunków w ms, oś pionowa dolnego rysunku w Hz, intensywniejsze zaczerwienie odpowiada większej amplitudzie chwilowej.

Obejrzyjmy teraz „mapę” pulsacji chwilowej pokazaną na rys. 5, na której w funkcji czasu wykreślono równocześnie przebiegi częstotliwości chwilowych (w Hz) dla wszystkich sygnałów wyjściowych zespolonego banku filtrów wąskopasmowych. Ten sposób jednoczesnego zobrazowania wszystkich pięciu reprezentacji AM-FM nazwaliśmy tutaj mapą, ponieważ dodatkowo, w postaci intensywności zaczerwienia, umieszczono na nim

informacje o wartości amplitudy chwilowej (im ciemniejszy punkt tym większa amplituda chwilowa). Już na podstawie tej mapy można powiedzieć, że w przypadku przetwarzanego sygnału mowy wybór częstotliwości chwilowej odpowiadającej sygnałowi o największej amplitudzie chwilowej, jako częstotliwości podstawowej tonu krtaniowego dla danej chwili czasu, jest najprostszym ale jednocześnie skutecznym rozwiązaniem. Można jednak równocześnie zauważyć, że w przypadku wyższych harmonicznych tonu krtaniowego odczytanie dokładnej wartości jest już utrudnione. Odstęp pomiędzy kolejnymi harmonicznymi jest stały, stąd też w przypadku wyższych harmonicznych często zachodzi sytuacja w której dwie kolejne składowe ułożone są w paśmie tego samego filtra. W efekcie przebieg częstotliwości chwilowej na wyjściu takiego filtra jest silnie pofalowany. Można to przykładowo zauważyć na rys. 5 dla drugiej harmonicznej tonu krtaniowego dla fragmentu do 1200ms do 1400ms.



Rys.6. Analizowany sygnał mowy (u góry) oraz (na dole) przebieg chwilowej wartości okresu tonu krtaniowego (w próbkach przy szybkości próbkowania 44 100 Sa/s): cienką czarną linią - proponowana metoda, szarą grubszą linią - metoda korelacyjna. Oś pozioma na wszystkich wykresach wyskalowana w ms.

Obejrzyjmy teraz „mapę” pulsacji chwilowej pokazaną na rys. 5, na której w funkcji czasu wykreślono równocześnie przebiegi częstotliwości chwilowych (w Hz) dla wszystkich sygnałów wyjściowych zespolonego banku filtrów wąskopasmowych. Ten sposób jednoczesnego zobrazowania wszystkich pięciu reprezentacji AM-FM nazwaliśmy tutaj mapą, ponieważ dodatkowo, w postaci intensywności zaczerwienia, umieszczono na nim informacje o wartości amplitudy chwilowej (im ciemniejszy punkt tym większa amplituda chwilowa). Już na podstawie tej mapy można powiedzieć, że w przypadku przetwarzanego sygnału mowy wybór częstotliwości chwilowej odpowiadającej sygnałowi o największej amplitudzie chwilowej, jako częstotliwości podstawowej tonu krtaniowego dla danej chwili czasu, jest najprostszym ale jednocześnie skutecznym rozwiązaniem. Można jednak równocześnie zauważyć, że w przypadku wyższych harmonicznych tonu krtaniowego odczytanie dokładnej wartości jest już utrudnione. Odstęp pomiędzy kolejnymi harmonicznymi jest stały, stąd też w przypadku wyższych harmonicznych często zachodzi sytuacja w której dwie kolejne składowe ułożone są w paśmie tego samego filtra. W efekcie przebieg częstotliwości chwilowej na wyjściu takiego filtra jest silnie pofalowany. Można to przykładowo zauważyć na rys. 5 dla drugiej harmonicznej tonu krtaniowego dla fragmentu do 1200ms do 1400ms.

4. WNIOSKI KOŃCOWE

W pracy zaprezentowano nową koncepcję działającego *on-line* estymatora tonu krtaniowego opartą o bank pięciu półoktawowych filtrów selektywnych pokrywających pasmo 50 – 283 Hz, typowe dla częstotliwości podstawowej tonu krtaniowego. Na wyjściu każdego z tych filtrów zastosowano proste demodulatory częstotliwości amplitudy obliczające reprezentację AM·FM sygnałów z poszczególnych podpasm, które dalej, w bloku decyzyjnym, wykorzystano do ostatecznego wyboru wartości chwilowej częstotliwości albo okresu tonu krtaniowego. Zaproponowany tu algorytm sprawdzono porównując obliczone przezeń estymaty tonu krtaniowego z estymatami tonu krtaniowego obliczonymi metodą korelacyjną, którą autorzy tego artykułu uznali za, jak dotychczas, najdokładniejszą.

BIBLIOGRAFIA

- [1] Kleijn W.B., Paliwal K.K. (Eds): *Speech Coding and Synthesis*. W: Elsevier 1995.
- [2] Seneff S.: *Real-time harmonic pitch detector*. W: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-26, nr 4, August 1978, pp. 358-365.
- [3] Rouat J., Yong Chun Liu, Morissette D.: *A pitch determination and voiced/unvoiced decision algorithm for noisy speech*. W: Speech Communication 21 (1997), pp. 191-207.
- [4] Medan Y., Yair E., Chazan D.: *Super resolution pitch determination of speech signals*. W: IEEE Transactions on Signal Processing, vol. 39, nr 1, January 1991, pp. 40-48.
- [5] Veprék P., Scordilis M.S.: *Analysis, enhancement and evaluation of five pitch determination techniques*. W: Speech Communication 37 (2002), pp. 249-270.
- [6] Atkinson I.A., Kondo A.M., Evans B.G.: *Pitch detection of speech signal using segmented autocorrelation*. W: Electronics Letters, vol. 31, nr 7, March 1995, pp. 533-535.
- [7] Rojewski M., Blok M., Blok E.: *Nowy estymator dyskretniej pulsacji chwilowej wykorzystujący jej dekompozycję na składowe: powolną i szybką*. W: Wojskowa Konferencja Telekomunikacji i Informatyki WKTiI-98, Zegrze, 7-9 października 1998, cz. 2 s. 217-226.
- [8] Mitra S.K., Kaiser J.F (Eds): *Handbook for Digital Signal Processing*. W: Wiley 1993.

A NEW ALGORITHM FOR PITCH PERIOD ESTIMATION

Summary

In this paper a concept of a new algorithm for on-line instantaneous pitch period estimation is presented. The proposed algorithm is based on a bank of narrowband complex filters, designed accordingly to the speech signal properties. Each filter in a bank is followed by a two-sample instantaneous frequency estimator. Firstly, in the algorithm each complex signal obtained at the output of the respective filter in the bank is demodulated in frequency and amplitude giving an AM·FM representation of the input. Next, on the basis of evaluated in this way AM·FM representations from all channels, the algorithm selects the frequency value, which corresponds to the instantaneous pitch frequency. The paper is illustrated with the results of experiments with speech signal processing using a computer implementation of the proposed algorithm.

**Andrzej Czyżewski*, Andrzej Kaczmarek*, Józef Kotus*,
Arkadiusz Pawlik**, Andrzej Rypulak**, Paweł Żwan***

***Katedra Systemów Multimedialnych, Politechnika Gdańska**

****Wydział Lotniczy, Wyższa Szkoła Oficerska Sił Powietrznych w Dęblinie**

CYFROWY SYSTEM REJESTRACJI I REKONSTRUKCJI SYGNAŁU MOWY DLA POTRZEB LOTNICTWA WOJSKOWEGO

Streszczenie

W pracy przedstawiono ogólną charakterystykę opracowanego systemu rejestracji i rekonstrukcji sygnału mowy. Zamieszczono skrótowy opis poszczególnych składników systemu, stanowiącego zestaw zaawansowanych narzędzi do rejestracji, analizy i rekonstruowania mowy, zrealizowany w formie oprogramowania komputerowego. Narzędzia te pozwalają na szybkie wyszukiwanie pożądaných fragmentów nagrań oraz poprawę ich jakości na drodze redukcji szumów, zniekształceń i zakłóceń. Przedstawiono również skróte informacje na temat wybranych algorytmów rekonstruowania mowy, których zastosowanie pozwoliło na uzyskanie szczególnie znaczącego przyrostu zrozumiałości przetwarzanej mowy.

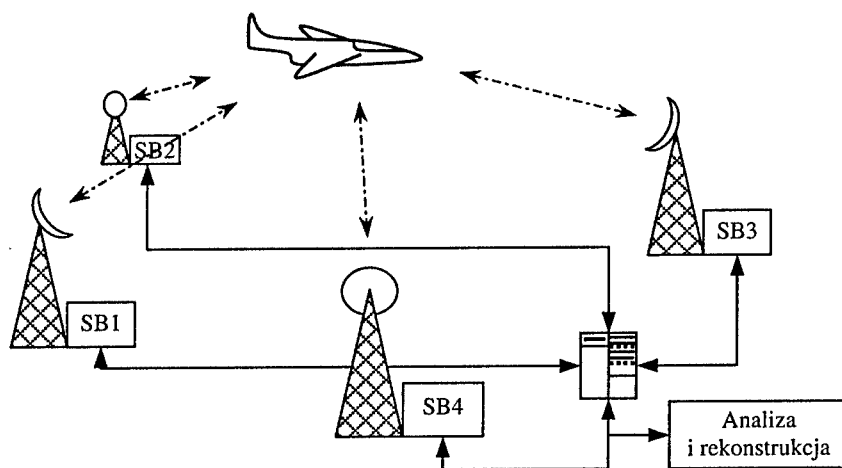
1. WPROWADZENIE

Komunikacja głosowa z pilotami samolotów lub śmigłowców jest obciążona poważnymi utrudnieniami, które wynikają z wpływu silnych zakłóceń akustycznych na sygnał mowy pilota oraz są skutkiem występowania zniekształceń, szumów i zaników powodowanych zmiennymi warunkami propagacji fal radiowych, wykorzystywanych do łączności pomiędzy samolotami i naziemnymi ośrodkami kontroli lotów. Ponadto, stosowane aktualnie metody akwizycji i analogowej rejestracji mowy pilotów w sposób istotny odbiegają od współczesnych wymagań technologicznych. Tymczasem, efektywna rejestracja mowy i poprawa jakości jej odbioru mają zasadnicze znaczenie dla zapewnienia bezpieczeństwa lotów oraz dla poprawy efektywności kształcenia pilotów i skuteczności prowadzonych analiz przebiegu lotów a także w procesie badania przyczyn katastrof lotniczych. W związku z powyższym, celem realizowanego projektu jest opracowanie i wprowadzenie do stosowania nowych rozwiązań technicznych i technologicznych w zakresie lotniczej komunikacji głosowej. W szczególności została opracowana metoda akwizycji i cyfrowej rejestracji sygnału mowy pilotów samolotów szkolno-treningowych w naziemnych ośrodkach kontroli lotów i zestaw narzędzi algorytmicznych, które pozwolą na efektywne wyszukiwanie pożądaných fragmentów nagrań oraz poprawę ich jakości na drodze redukcji szumów, zakłóceń i zniekształceń. Cyfrowy system rejestracji nagrań umożliwi efektywną

kompresję mowy, która z kolei, czyni możliwym wielogodzinny zapis komputerowy przebiegu komunikacji głosowej i jego archiwizację. Jednocześnie, dzięki specjalnie opracowanej metodzie opartej na rozszerzonym kodowaniu perceptualnym mowy, w procesie kompresji eliminowane są szum i inne zakłócenia utrudniające odbiór i rozumienie mowy pilotów. W skrajnie trudnych lub szczególnie istotnych przypadkach analizy zapisu, możliwe jest wykorzystanie większej liczby (do pięciu) wersji tego samego nagrania zarejestrowanego w poszczególnych naziemnych stacjach kontroli lotu, dzięki specjalnie opracowanej metodzie synchronizacji zapisu i jego odtwarzania w fonicznym zestawie wielokanałowym. Odtwarzanie wielokanałowe jest stosowane w celu radykalnej poprawy warunków odsłuchu, która stanie się możliwa dzięki wykorzystaniu subiektywnej lokalizacji głosu pilota w innym punkcie przestrzeni akustycznej, niż lokalizacja pozornych źródeł dźwięku pochodzących od obecności nieskorelowanych zakłóceń rejestrowanych wraz z sygnałem mowy w poszczególnych stacjach naziemnych. Metoda ta umożliwia wtórną poprawę warunków odbioru radiowego w sytuacji dysponowania kilkoma wersjami zapisu z nasłuchu radiowego. Zapisany komputerowo sygnał mowy może być ponadto synchronizowany z innymi systemami odtwarzania przebiegu lotu. W dalszej części pracy zostaną przedyskutowane założenia i realizacja poszczególnych ogniw składowych opracowanego systemu do rejestracji i rekonstruowania zapisu komunikacji głosowej z pilotami wojskowych statków powietrznych.

2. OGÓLNA CHARAKTERYSTYKA SYSTEMU

Ogólną strukturę systemu można przedstawić jako zbiór stacji bazowych (rejestratorów) dokonujących rejestracji korespondencji radiowej oraz komunikujących się za pośrednictwem serwera. Jest to rozproszony system rejestrujący korespondencję radiową, prowadzoną na określonym obszarze. Ogólny schemat systemu przedstawiono na rysunku 1.



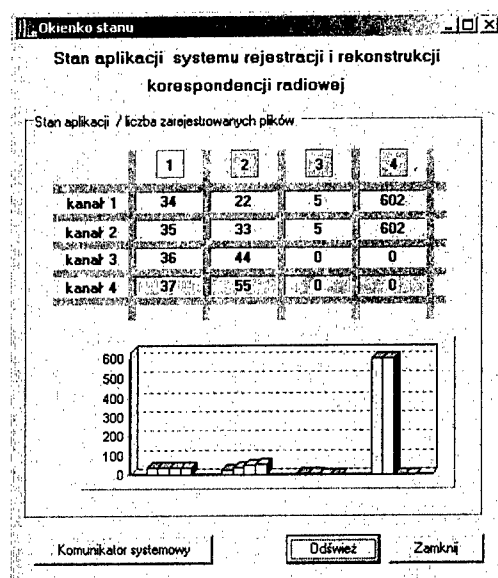
Rys.1. Ogólny schemat systemu rejestracji korespondencji radiowej

Opracowany system umożliwia rejestrację, która jest prowadzona w określonych stacjach bazowych. Zarejestrowany materiał stanowi materiał źródłowy, wykorzystywany

dalej w procesie analizy. Do poprawnego działania systemu niezbędne było stworzenie odpowiedniego oprogramowania. W toku realizacji projektu wyłoniła się rzeczywista struktura informatyczna obejmująca zbiór programów realizujących określone zadania. Do zapewnienia nieprzerwanego monitorowania kanałów komunikacyjnych oraz rejestrowania korespondencji stworzono program **Rejestrator**. Aplikacja ta gromadzi ponadto zarejestrowane dane oraz umożliwia przesyłanie do serwera. Do analizy zarejestrowanej korespondencji z zadanego okresu czasu opracowano program pod nazwą **Przeglądarka**. Umożliwia on pobranie z serwera zapisanej korespondencji, ponadto udostępnia mechanizmy wyszukiwania korespondencji spełniające zadane kryteria. Przeglądarka umożliwia szybką analizę sesji nagranych. Podczas analizy korespondencji silnie zakłóconej niezbędna jest dodatkowa obróbka sygnału, zmierzająca do poprawy zrozumiałości mowy. Do tego celu stworzono wyspecjalizowane narzędzie, udostępniające pakiet zaawansowanych algorytmów z dziedziny cyfrowego przetwarzania sygnałów. Aplikację tę nazwano **Rekonstruktor** – umożliwia ona rekonstruowanie zniekształconego sygnału mowy. Opracowano ponadto dedykowaną bazę danych na serwerze, przechowującą archiwalną korespondencję. Do komunikacji sieciowej pomiędzy poszczególnymi aplikacjami stworzono specjalny protokół komunikacyjny. Umożliwia on, oprócz przesyłania komunikatów sterujących i informacyjnych pomiędzy stacjami bazowymi, komunikowanie się między sobą osób obsługujących stacje bazowe.

3. REJESTRACJA KORESPONDENCJI RADIOWEJ

Aplikacja rejestratora jest programem realizującym kompleksowe zadanie rejestracji korespondencji radiowej. Umożliwia wizualizację tego procesu w czasie rzeczywistym. Zapewnia również obsługę zarejestrowanej korespondencji, jej przechowywanie na lokalnym dysku oraz archiwizowanie na centralnym serwerze. Program jest uruchamiany na komputerze klasy PC z zainstalowanym systemem operacyjnym Windows 2000, wyposażonym w wielokanałową kartę dźwiękową. Stanowi podstawowe ogniwo opracowanego systemu. Jest zaprojektowany z myślą o pracy sieciowej realizując w ten sposób funkcję wielokanałowego rozproszonego rejestratora. Dla zapewnienia wzajemnej komunikacji pomiędzy stacjami bazowymi opracowano odpowiedni protokół komunikacji sieciowej. Komunikacja pomiędzy poszczególnymi stacjami bazowymi odbywa się za pośrednictwem serwera. W momencie uruchomienia aplikacji Rejestrator, nawiązywane jest połączenie sieciowe programu z serwerem. Jeśli komunikacja zostanie nawiązana to program Rejestrator loguje się do serwera. Po stronie serwera jest tworzony dziennik wszystkich odebranych i wysłanych komunikatów z informacją o czasie wysłania i odebrania. Po poprawnym zalogowaniu się Rejestrator wykonuje operację synchronizacji zegara systemowego z zegarem serwera. W ten sposób jest realizowana wzajemna synchronizacja poszczególnych stacji bazowych, gdyż mają one czas zgodny z zegarem serwera. W następnej kolejności jest przekazywana informacja o stanie aplikacji, którą z kolei serwer przekazuje do wszystkich aktualnie zalogowanych stacji bazowych. Pobierany jest również stan wszystkich aktualnie zalogowanych stacji bazowych, tzn. czy jest dokonywana rejestracja oraz ile korespondencji zarejestrowano w poszczególnych kanałach komunikacyjnych. Po zakończeniu wstępnej inicjalizacji program Rejestrator jest gotowy do rozpoczęcia rejestracji. Po jej uruchomieniu automatycznie zostaje wysłany odpowiedni komunikat, który jest kierowany przez serwer do pozostałych stacji bazowych. Na rysunku 2 przedstawiono wygląd okna dialogowego prezentującego stan poszczególnych stacji bazowych, odczytany na podstawie przesyłanych komunikatów.



Rys.2. Okno komunikacji sieciowej

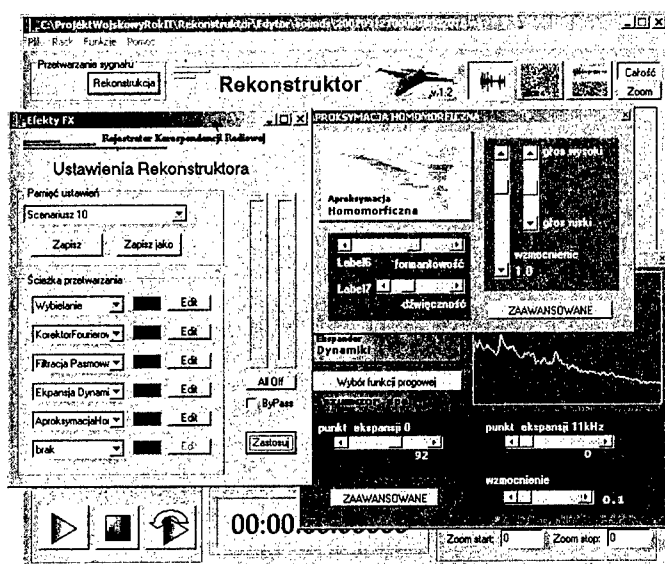
4. ANALIZA ZAREJESTROWANEJ KORESPONDENCJI

Do przeprowadzenia analizy zarejestrowanej korespondencji radiowej opracowano aplikację pod nazwą Przeglądarka. Jej główne funkcje to: pobieranie z serwera plików z sesji nagraniowej z zadanego dnia lub spełniających zadane kryteria (np. nr kanału rejestracyjnego, nr stacji bazowej, konkretny przedział czasu), przedstawianie ich na skali czasu, odsłuch korespondencji w systemie wielokanałowym. Aplikacja ta posiada ponadto szereg funkcji usprawniających odtwarzanie dźwięku, należą do nich: wzajemna synchronizacja wypowiedzi pochodzących z różnych stacji bazowych, normalizacja głośności dźwięku odsłuchiwanym wypowiedzi, konfiguracja odsłuchu wielokanałowego oraz współpraca z programem Rekonstruktor, służącym do rekonstrukcji mowy.

5. REKONSTRUKCJA SYGNAŁU MOWY

Proces rekonstruowania pliku silnie zniekształconego, o niskiej zrozumiałości, jest zagadnieniem złożonym i czasochłonnym. Wymaga gruntownej znajomości z zakresu wytwarzania i percepcji mowy jak również umiejętności w posługiwaniu się algorytmami cyfrowego przetwarzania sygnałów. Autorzy programu dołożyli wszelkich starań, by ten proces maksymalnie uprościć, zwracając przy tym uwagę, by zachować maksimum skuteczności w działaniu poszczególnych funkcji. Program Rekonstruktor charakteryzuje unikatowy zbiór różnorodnych algorytmów, przydatnych w procesie wydobywania treści przekazu z silnie zniekształconych plików. Do dyspozycji użytkownika oddano szereg mechanizmów ułatwiających używanie poszczególnych funkcji. Zrealizowano to przez podział algorytmów, ze względu na sposób dokonywania operacji na sygnale, na dwie grupy. Pierwszą grupę stanowią algorytmy działające w czasie rzeczywistym (**algorytmy działające w trybie on-line**). Ich działanie polega na tym, że użytkownik, po uruchomieniu

mieniu wybranego algorytmu należącego do tej grupy, może, podczas odtwarzania przetwarzanego pliku, na bieżąco słyszeć wynik przetwarzania w zależności od wartości ustawionych parametrów. Może również dokonywać dowolnych manipulacji na parametrach i natychmiast usłyszeć, jakie wyniki dają te zmiany. W praktyce sprawdzono, że takie podejście znacząco ułatwia wyszukiwanie optymalnych ustawień algorytmu do konkretnego przykładu dźwiękowego. Do obsługi algorytmów z tej grupy opracowano specjalny interfejs, pokazany na rys. 3. Umożliwia on szeregowe podłączenie do ośmiu równocześnie działających algorytmów, dostępnych z wybranej listy rozwijanej.



Rys.3. Interfejs do obsługi algorytmów w trybie on-line

Do drugiej grupy należą procedury, których działanie wymaga jednoczesnej obróbki całego sygnału, z tego też powodu nie można kontrolować słuchowo efektu ich działania i jednocześnie zmieniać wartości wybranych parametrów. Tę grupę nazwano **algorytmami działającymi w trybie off-line**. Działanie algorytmów z drugiej grupy polega na ich wybraniu, dokonaniu ustawień parametrów (w niektórych przypadkach możliwe jest wstępne odsłuchanie wyniku) i wciśnięciu klawisza Zastosuj. W następstwie tej akcji powstaje przetworzony plik, którego przebieg czasowy pojawia się w głównym oknie. W tab. 1 zestawiono wszystkie algorytmy dostępne w programie Rekonstruktor dla poszczególnych grup.

Ustalono doświadczalnie, że proces rekonstrukcji można znacząco przyspieszyć i ułatwić, poprzez zapisywanie ustawień poszczególnych algorytmów, ich wzajemnych kombinacji (**scenariusz rekonstrukcji**) lub przez zapisanie całego procesu przetwarzania (**scenariusz przetwarzania**). Ułatwienia te polegają na tym, że możliwe jest późniejsze przywołanie wcześniej ustalonych nastaw do różnych plików charakteryzujących się np. podobieństwem zniekształceń występujących w sygnale.

Tablica 1

Zestawienie algorytmów dostępnych w programie Rekonstruktor

Algorytmy działające on-line	Algorytmy działające off-line
Korektor pasm mowy	Pasmowy procesor dynamiki
Bramka szumów	Eliminator przesterowań
Ogranicznik poziomu sygnału	Analizator widmowy
Reduktor szumów	Moduł transpozycji czasu
Moduł wybielania szumu	Korektor neuronowy
Moduł ślepego rozplatania	Dekorelator
Ekspander dynamiki widma	
Korektor fourierowski	
Moduł aproksymacji zespolonej	
Moduł aproksymacji homomorficznej	

6. WYBRANE ALGORYTMY REKONSTRUKCJI SYGNAŁU MOWY

6.1 Moduł transpozycji czasu

Celem tego modułu jest wydłużenie czasu trwania nagrania bez zmiany wysokości głosu, czyli ogólnie: wszystkich częstotliwości składowych (tonu krtaniowego). Algorytm polega na powieleniu automatycznie selekcjonowanych fragmentów sygnału mowy (pseudookresów tonu krtaniowego), przez co uzyskuje się dłuższy czas odtwarzania. Wydłużenie jest możliwe w zakresie 0-100%. Maksymalna wartość 100% oznacza podwojenie czasu trwania całego nagrania. Do dyspozycji są dwa sposoby wydłużenia opierające się na dwóch różnych metodach zakładkowania: zakładkowaniu międzybuforowym i zakładkowaniu wewnętrznym. Dodatkowo została zaimplementowana analiza cepstralna, przy pomocy której możliwa jest estymacja częstotliwości tonu krtaniowego, co zapewnia lepsze zachowanie ciągłości fazy i minimalizację szkodliwych skutków segmentacji.

Do detekcji tonu krtaniowego zastosowana została analiza cepstralna. Obliczane jest chwilowe cepstrum mocy sygnału na podstawie widma sygnału zawartego w pojedynczym buforze zawierającym 1024 próbki. Realizowana jest jednoczesna detekcja tonu krtaniowego i estymacja jego częstotliwości. Wyniki te są uzyskiwane przy pomocy algorytmu poszukującego lokalnego maksimum w uzyskanym cepstrum. Poszukiwania są ograniczone do zakresu częstotliwości od 78 Hz do 719 Hz (około), co obejmuje częstotliwości tonu krtaniowego większości głosów. Dodatkowo wprowadzone jest ograniczenie zadawane przez użytkownika, wynikające ze znajomości zakresu zmian częstotliwości tonu krtaniowego dla konkretnej wypowiedzi. Wpływa to na uniknięcie pomyłek podczas detekcji tonu krtaniowego w przypadku sygnału silnie zakłóconego. Ponadto wprowadzono mechanizm śledzenia maksimum cepstrum poprzez ważone okienkowanie współczynników cepstralnych, które jest wykonywane na podstawie zapamiętanego położenia maksimum z poprzedniego segmentu. W przypadku segmentów bezdźwięcznych wykryte maksimum jest natomiast ignorowane na podstawie zadawanego przez użytkownika odpowiedniego progu detekcji. Zostaje w tym przypadku zachowane maksimum z ostatniego segmentu dźwięcznego. W efekcie detektor tonu krtaniowego zachowuje ciągłość wyników, pomimo obecności zakłóceń, maskujących sygnał mowy. Wykryte maksimum służy do estymacji

częstotliwości tonu krztaniowego. Do tych obliczeń wykorzystuje się dodatkowe informacje o bezpośrednim jego sąsiedztwie i stosuje się interpolację kwadratową w celu pokonania ograniczeń wynikających z rozdzielczości analizy cepstralnej.

6.2. Ekspander dynamiki widma

Poprawa wyrazistości mowy następuje w tym przypadku poprzez zwiększenie odstepu poziomu sygnału użytecznego od poziomu szumu i zakłóceń. Założeniem dla tej metody jest stacjonarność sygnału zakłócającego. Algorytm jest realizowany w dziedzinie widma w pasmach wybranych na podstawie porównania poziomów: obserwowanego (zakłócony sygnał mowy) i referencyjnego (wzorcowy sygnał szumu).

Idea wykorzystania średniego widma szumu i zakłóceń jest podobna do idei będącej podstawą dla typowego algorytmu odejmowania widma. Różnica polega na tym, że wykorzystywane widmo średnie traktuje się jako element pomocniczy podczas tworzenia funkcji progowej ustalającej zakresy częstotliwości, podlegające ekspansji dynamiki. W związku z tym wprowadzono dwupunktową regulację charakterystyki ekspansji związaną ze skrajnymi częstotliwościami przetwarzanego pasma: 0 i 11025 Hz. Można w ten sposób wpływać na wielkość ingerencji w sygnale obierając jako kryterium wyrazistość i zrozumiałość mowy.

Zakłada się następnie, że sygnał użyteczny dominuje nad zakłóceniami i możliwe jest ustalenie poziomu (wielkości progowej) o wartości pośredniej, który pozwoli na redukcję tych zakłóceń. Procedura ekspansji dynamiki została opracowana dla widma mocy liczonego przy pomocy algorytmu FFT. Powyżej zadanego progu nie odbywa się żadna modyfikacja sygnału, zaś poniżej tego progu odbywa się przekształcenie realizujące ekspansję dynamiki poprzez przemnożenie zespolonych składowych transformaty Fouriera przez współczynniki proporcjonalne do wartości odpowiadających im modułom. Powoduje to zwiększenie dynamiki z zachowaniem ciągłości charakterystyki.

Zerowanie widma poniżej progu powodowałoby nieciągłość charakterystyki (byłoby to odejmowanie widma), natomiast zastosowana procedura zapewnia tę ciągłość a charakterystyka ta daje się opisać w postaci linii łamanej. Zastosowanie funkcji drugiego rzędu powoduje podwojenie dynamiki (np. komponenty widma będące pierwotnie w odstepie 6 dB po ekspansji znajdują się w odstepie 12 dB). Próg ekspansji może być różny dla dolnej i górnej części pasma, można także dobierać kształt funkcji progowej dla wszystkich punktów całego pasma stosownie do właściwości widmowych zakłóceń.

6.3. Moduł aproksymacji zespolonej

Poprawa wyrazistości mowy poprzez wygładzenie widma zespolonego. Oczekuje się w tym przypadku lepszego uwypuklenia głosek zwartych (plozywnych) i dźwięcznych w stosunku do zakłóceń szumowych.

Zakłada się, że w krótkim przedziale czasu (wybrano wartość ok. 46ms) sygnał mowy jest sygnałem deterministycznym, który charakteryzuje się wykresem fazowym bardziej gładkim, aniżeli identyczny wykres dla sygnału stochastycznego. Dla osiągnięcia podobieństwa widma fazowego sygnału zakłóconego do sygnału nagranych w dobrych warunkach stosuje się operację wygładzania na widmie zespolonym. Korzysta się z sąsiedztwa czterech zespolonych punktów widma (po dwa z każdej strony), licząc jako wynik kombinację liniową wartości tych punktów, co stwarza możliwość objęcia zakresu od aproksymacji wielomianowej rzędu 3 poprzez średnią arytmetyczną dwóch punktów do

ekstrapolacji liniowej wprzód i wstecz. W ostatecznym rozwiązaniu aproksymacja sprowadza się do obliczenia kombinacji liniowej sąsiednich czterech punktów widma (po dwa z każdej strony). Dla uproszczenia można założyć numerację punktów widma: -2, -1, 0, 1 i 2, przy czym punkt 0 jest punktem aproksymowanym. Wartości widma odpowiadające tym punktom to odpowiednio: x_{-2} , x_{-1} , x_0 , x_1 , i x_2 . Ostatecznie aproksymowany punkt x_0 wyraża się wzorem 1:

$$\hat{x}_0 = \frac{1}{24} \cdot [w_w \cdot (x_{-1} + x_1) - (w_w - 12) \cdot (x_{-2} + x_2)] \quad (1)$$

gdzie: w_w – współczynnik wygładzania z przedziału $\langle 0, 24 \rangle$

Aproksymacja wielomianowa rzędu 3 jest wykonywana dla wartości $w_w = 16$. Inne sposoby wygładzania to np. średnia arytmetyczna dwóch punktów (dwa przypadki: sąsiednich lub oddalonych o 1 – odpowiednio $w_w = 12$ i $w_w = 0$), ekstrapolacja liniowa wprzód i wstecz ($w_w = 24$).

Lokalne maksima widma w sygnale mowy są związane ze składowymi harmonicznymi tonu krtaniowego. Ze względu na własności tonu krtaniowego (obecność wszystkich składowych harmonicznnych) możliwe staje się zastąpienie śledzenia maksimów widma poprzez śledzenie maksimów cepstrum. Śledzenie to jest związane z ograniczonym pasmem i odpowiednią funkcją wagową, wskutek tego zmniejsza się prawdopodobieństwo pomyłki podczas estymacji częstotliwości tonu krtaniowego.

7. ZAKOŃCZENIE

Z uwagi na ograniczoną objętość tekstu, możliwe było jedynie skrótowe przedstawienie struktury opracowanego systemu i tylko wybranych algorytmów, które są w jego ramach stosowane do przetwarzania sygnału mowy [1]. Jak wskazują wstępne wyniki eksploatacji systemu w Wyższej Szkole Sił Powietrznych w Dęblinie, opracowany system spełnia oczekiwania użytkowników, pozwalając na znaczącą poprawę zrozumiałości zapisu komunikacji głosowej z pilotami wojskowych statków powietrznych.

BIBLIOGRAFIA

- [1] Dokumentacja projektu celowego nr 148346/C-T00/2002. Katedra Systemów Multimedialnych. Wydział ETI. Politechnika Gdańska, 2004 r.

SPEECH ARCHIVING & RESTORATION SYSTEM FOR MILITARY AVIATION APPLICATIONS

Summary

The speech received by radio communication from jet pilots can be severely degraded by noise and various distortions. A system was developed for multi-channel recording of voice communication with jet pilots extended with a toolbox containing some advanced DSP algorithms for speech enhancement. Moreover, some innovative solutions were adopted, including the method for synchronizing transmission received from many radio stations in order to produce surround sound enabling additional perceptual filtration of speech. Some selected components of the engineered multi-task speech enhancement system are presented in the paper.

Ewa Hermanowicz^{*}, Mirosław Rojewski^{**}

^{*} Katedra Systemów Multimedialnych, ^{**} Katedra Systemów Informacyjnych
Politechnika Gdańska

KWADRATUROWY DDS Z UŁAMKOWO-OPÓŹNIAJĄCYM FILTREM O STRUKTURZE FLASH-FARROW

Streszczenie

W pracy omawiamy nieliniowy algorytm cyfrowego generatora z synteza bezpośrednią, zwanego krótko DDS od ang. *Direct Digital Synthesizer*. Proponujemy nowy algorytm kwadraturowego DDS. Pozwala on osiągnąć zarówno wysoki stopień czystości generowanej sinusoidy kwadraturowej, jak i bardzo małe błędy modulacji częstotliwości (FM) przy małej pojemności pamięci ROM, od której zależy pobór mocy zasilania. Ponadto proponujemy nowe podejście do ilościowej oceny jakości DDS, odpowiednie do określania nie tylko czystości generowanej zespolonej sinusoidy, ale również błędów modulacji.

1. WSTĘP

Jednym z podstawowych algorytmów stosowanych na poziomie sygnałowym wielu współczesnych technologii informacyjnych jest DDS (ang. *Direct Digital Synthesizer*). Jest to działający potokowo (ang. *on-line*) algorytm cyfrowego przetwarzania sygnałów (CPS). Jego definicyjnym zadaniem jest cyfrowa synteza czystej niemodulowanej sinusoidy rzeczywistej albo zespolonej. Od kilkunastu lat DDS do swej pierwotnej roli oscylatora cyfrowego dołączył rolę uniwersalnego modulatora, przede wszystkim częstotliwości (FM), ale też pośrednio fazy (PM) i amplitudy (AM), przebiegu sinusoidalnego. Dziś DDS jako oscylator bądź modulator znajduje liczne zastosowania w różnorodnych systemach i urządzeniach telekomunikacji, elektroniki profesjonalnej i użytkowej, inżynierii multimedialnej i nawigacyjnej, geofizyki i astrofizyki, chronometrii, metrologii i innych.

Algorytm DDS może być implementowany na każdym procesorze – a więc jako program komputerowy, na zmiennie- albo stałoprzecinkowym procesorze DSP (ang. *Digital Signal Processor*), w programowalnym układzie cyfrowym lub jako jeden z algorytmów układu scalonego ASIC (ang. *Application Specific Integrated Circuit*). Wymagania stawiane dzisiejszym DDSom zależą, rzecz jasna, od zastosowań i są dwojakiego rodzaju. Po pierwsze, wierność i dokładność generowanego sygnału: czystość oscylacji, modulacja bez zniekształceń. Po drugie, jednoczesna miniaturyzacja hardware'owa i oszczędność poboru mocy zasilania. I jedno i drugie stawia przed inżynierami CPS nowe wyzwanie. Jest nim

opracowanie nowych algorytmów DDS bazujących na coraz to mniejszej pojemności tablicy trygonometrycznej zwanej krótko LUT od ang. *look-up table*, lub tablicy wyników pośrednich algorytmu oscylatora, przechowywanych w pamięci ROM układu scalonego. Dotyczy to w szczególności zastosowań w urządzeniach przenośnych takich, przykładowo, jak telefon komórkowy, odbiornik GPS (ang. *Global Positioning System*), czy kieszonkowy odbiornik radiowy.

Do dnia dzisiejszego znanych jest zaledwie parę podejść do zagadnienia minimalizacji wymiaru LUT zapisanej w ROM. Wszystkie je cechuje ta sama właściwość: rozrzedzeniu LUT nierozłącznie towarzyszy wzrost złożoności obliczeniowej algorytmu DDS. W rozwiązaniach katalogowych DDSa obserwuje się dwie skrajności. Pierwsza to rozwiązanie najbardziej obliczeniowo oszczędne, ale za to o kompletnej LUT dużego rozmiaru, z której można odczytać wartości $\cos \varphi$ lub $\sin \varphi$ dla zbioru wszystkich faz chwilowych syntetyzowanego przebiegu. W tym rozwiązaniu spotykamy się z LUT nawet o 2^B komórkach [1], gdzie B oznacza liczbę bitów zakładanego kroku fazy φ . Drugą skrajność stanowią DDSy nie posiadające LUT (ROM-less DDS) [2]. W tym rozwiązaniu wszystkie wartości próbek syntetyzowanego przebiegu oblicza się numerycznie angażując wielokrotnie większe moce obliczeniowe niż w pierwszym przypadku. Ale przedmiotem większości ostatnich publikacji są rozwiązania pośrednie, omawiane np. w [1], [3], z pamięcią (z reguły) od kilkunastu do kilkudziesięciu równoodległych próbek jednego okresu sinusoidy w LUT. Do zagęszczania LUT stosuje się wtedy głównie klasyczne metody interpolacji prawie wyłącznie linearnej (poprawki proporcjonalne) [3].

W tej pracy zajmujemy się rozwiązaniem DDSa również należącym do klasy pośredniej. Jego zaletą jest, iż pozwala ono zbliżyć się do potencjalnej granicy dokładności DDSa, określonej przez symulacyjne badanie DDSa bez kwantyzacji. Drogą do tego celu jest nowe podejście do DDSa jako do zespolonego algorytmu próbkowania (resamplera) zespolonej sinusoidy dyskretniej, zapisanej w LUT, na sinusoidę wyjściową niemodulowaną lub modulowaną FM. Narzędziem do tego celu jest zespolony filtr ułamkowo-opóźniający (CFDF – ang. *complex fractional-delay filter*) [4] dotychczas w DDS nie stosowany. DDS, potraktowany w p. 2 jako zespolony resampler zmiennoprzecinkowy, generuje tak dokładne przebiegi, że bezpośrednie ich porównywanie z odpowiednikami obliczonymi ze wzorów analitycznych za pomocą MATLABa wypada na niekorzyść MATLABa. Pojawiła się potrzeba opracowania nowego kryterium czystości zespolonego przebiegu generowanego przez DDS, o zadanej częstotliwości chwilowej i amplitudzie chwilowej. Kryterium to, nazwane przez nas zespolonym błędem chwilowym, przedstawiamy w p. 3, gdzie również oceniamy wyniki pracy naszego DDSa jako QO (ang. *quadrature oscillator*). W końcowym p. 4 podsumowujemy tę pracę i wyciągamy wnioski.

2. OPIS NOWEJ KONCEPCJI DDS

Zespolony/kwadraturowy DDS wg naszej koncepcji jest kaskadą trzech algorytmów nieliniowego CPS, jak przedstawiono na rys.1. Są to:

- 1) zwijający akumulator fazy (ang. *phase accumulator* – PA),
- 2) kwantyzator fazy (ang. *phase quantizer* – PQ),
- 3) resampler ciągu zespolonego (ang. *complex resampler* – CR).

Ciągiem wejściowym PA, a więc i DDSa jako całości, jest przebieg pulsacji chwilowej w radianach na próbkę [rad/Sa]

$$\omega[n] \in (-\pi, \pi) \quad \forall n = 0, 1, 2, \dots \quad (2.1)$$

gdzie n jest numerem bieżącej próbki w DDSie, przy czym $n = 0$ oznacza chwilę początkową – chwilę uruchomienia tego algorytmu. DDS startuje z fazy początkowej $\phi_0 \in (-\pi, \pi)$, która może być przypadkowa albo zadana z góry. Gdy pulsacja chwilowa jest stała

$$\omega[n] = \omega_c = \text{const}[n], \quad |\omega_c| \in (0, \pi) \quad (2.2)$$

to proponowany DDS pełni rolę QO, którego zadaniem jest generacja sinusoidy zespolonej

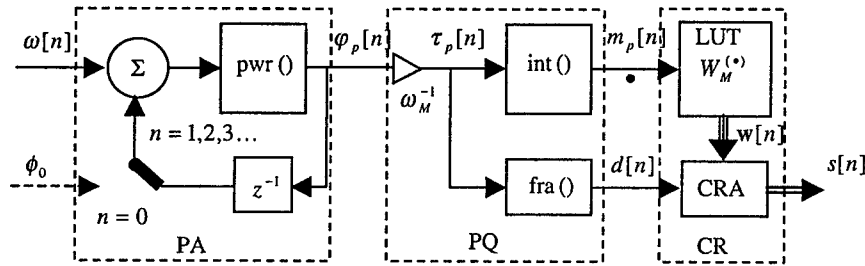
$$\exp(j(\phi_0 + \omega_c n)) \quad n = 0, 1, 2, \dots \quad (2.3)$$

W przeciwnym przypadku, a więc gdy pulsacja chwilowa (2.1) jest zależna od czasu, DDS jest modulatorem FM, a jego przebieg wejściowy $\omega[n]$ jest ciągiem modulującym sygnał FM. Kwadraturowy DDS, zarówno jako QO albo jako modulator FM, powinien generować przebieg zespolony

$$\exp(j(\phi[n])) \quad n = 0, 1, 2, \dots \quad (2.4)$$

którego faza chwilowa $\phi[n]$ stanowi wynik akumulacji zadanej pulsacji chwilowej $\omega[n]$ przy fazie początkowej ϕ_0 , czyli

$$\phi[n] = \begin{cases} \phi_0, & n = 0 \\ \phi_0 + \sum_{k=1}^n \omega[k], & n = 1, 2, 3, \dots \end{cases} \quad (2.5)$$



Rys.1. Schemat blokowy proponowanego DDSa. Oznaczenia w tekście.

W powyższym wzorze, ze wzrostem n , faza chwilowa $\phi[n]$ może zmierzać do $\pm\infty$. O takiej fazie mówimy, że jest rozwinięta (ang. *unwrapped*) poza okres jej wartości głównych $[-\pi, \pi)$. Występowanie fazy rozwiniętej (2.5) w teoretycznym zapisie (2.4) jest rzeczą ogólnie przyjętą. Natomiast w obliczeniach numerycznych, ze względu na okresowość funkcji $\exp(\cdot)$ dla argumentu urojonego, należy ten argument, tj. $j\phi[n]$, przeliczyć (zwinąć, ang. *wrap*) do okresu głównego (ang. *principal period*, stąd indeks dolny p w dalszych oznaczeniach). Wartość główną (ang. *principal value*) $\phi_p[n]$ obliczamy (rys.1) za pomocą akumulatora fazy PA zwinającego ją do okresu głównego. Zapisujemy to jako

$$\phi_p[n] = \begin{cases} \phi_0, & n = 0 \\ \text{pwr}(\phi_p[n-1] + \omega[n]), & n = 1, 2, 3, \dots \end{cases} \quad (2.6)$$

gdzie $\phi_p[n] \in [-\pi, \pi)$. Funkcja $\text{pwr}(\cdot)$ znana jest pod nazwą zwinacza fazy (ang. *phase wrapper*). Jej definicja jest następująca

$$\phi_p = \text{pwr}(\alpha) := 2\pi \text{fra}(\alpha/(2\pi)) \quad \forall \alpha \in \mathbf{R} \quad (2.7)$$

gdzie \mathbf{R} oznacza zbiór liczb rzeczywistych, a

$$\text{fra}(x) := x - \lfloor x + 1/2 \rfloor \in [-1/2, 1/2) \quad (2.8)$$

oraz

$$\text{int}(x) := \lfloor x + 1/2 \rfloor = x - \text{fra}(x) \quad (2.9)$$

oznaczają część ułamkową i część całkowitą liczby rzeczywistej x . $\lfloor x \rfloor$ oznacza największą liczbę całkowitą nie większą od x (w MATLABie funkcja `floor()`). Obliczanie, w każdej chwili n , wartości głównej $\varphi_p[n]$ wg algorytmu PA jest pierwszym zadaniem DDSa.

Drugim i zarazem głównym zadaniem DDSa jest znajdowanie wartości chwilowych (próbek) zespolonego ciągu wyjściowego, którym ma być

$$\exp(j(\varphi_p[n])); \quad n = 0, 1, 2, \dots \quad (2.10)$$

Można go otrzymać, jak wspominaliśmy we wprowadzeniu, na trzy różne sposoby: obliczyć (2.10) numerycznie albo odczytać z tzw. kompletnej LUT, zapamiętanej w ROM, albo na sposób pośredni, oparty o LUT zawierającą M próbek jednego okresu zespolonej sinusoidy

$$W_M^m := \exp(j2\pi m/M); \quad m = -\lfloor M/2 \rfloor, \dots, -1, 0, 1, \dots, \lfloor (M-1)/2 \rfloor \quad (2.11)$$

Liczbę m nazywamy numerem rubryki LUT albo adresem fazowym. Obliczone w PA fazy chwilowe $\varphi_p[n]$ z zasady przyjmują wartości różne od faz $2\pi m/M$ reprezentowanych w tablicy (2.11). Naturalnym i przeważnie wykorzystywanym sposobem oszacowania próbki zespolonej (2.10) jest interpolacja linearna pomiędzy dwoma sąsiednimi próbkami tablicy (2.11), pomiędzy fazami których (powiedzmy: $2\pi m_1/M$ i $2\pi(m_1+1)/M$) lokuje się faza $\varphi_p[n]$ [3], [5]. Można, rzecz jasna, pomyśleć o bardziej skomplikowanych, a dokładniejszych, lecz dłuższych interpolatorach LUT. W praktyce są jednak stosowane interpolatory krótkie, co najwyżej drugiego rzędu (np. [2]), ze względu na większe koszty numeryczne bardziej zaawansowanego interpolatora, rosnące z kwadratem rzędu.

W proponowanej niżej naszej koncepcji DDSa staramy się pogodzić mały wymiar LUT z małą złożonością obliczeniową algorytmu. Udać się to osiągnąć dzięki zupełnie innemu podejściu do opisanego powyżej zagadnienia obliczania próbki $\exp(j(\varphi_p[n]))$, gdy faza $\varphi_p[n]$ różni się od każdej z faz $2\pi m/M$ reprezentowanych w tablicy (2.11). W naszym podejściu do QO obliczanie ciągu wyjściowego DDSa utożsamiamy z przepróbkowaniem zespolonym zapamiętanej w LUT (2.11) zespolonej sinusoidy

$$\exp(j\omega_M m/M) \quad (2.12)$$

o stałej pulsacji

$$\omega_M := 2\pi/M \quad (2.13)$$

na, w ogólności modulowaną, sinusoidę zespoloną (2.10) o zadanej pulsacji chwilowej (2.1), a więc o ciągu faz chwilowych $\varphi_p[n]; n = 0, 1, 2, \dots$. W tym celu w każdym taktie pracy DDSa, czyli dla każdej chwili n , należy przeliczyć fazę chwilową $\varphi_p[n]$ na opóźnienie fazowe $\tau_p[n]$ obliczanej próbki wyjściowej na osi „czasu” m przebiegu (2.12), (2.13). Korzystamy tu z definicji opóźnienia fazowego

$$\tau_p[n] := \varphi_p[n]/\omega_M \in [-\lfloor M/2 \rfloor, \lfloor (M-1)/2 \rfloor] \quad (2.14)$$

a następnie rozkładamy $\tau_p[n]$ na części: całkowitą i ułamkową, wg wcześniej wprowadzonych definicji, odpowiednio, (2.9) i (2.8):

$$\tau_p[n] = m_p[n] + d[n] \quad (2.15)$$

gdzie

$$m_p[n] := \text{int}(\tau_p[n]) \in \{-\lfloor M/2 \rfloor, \dots, -1, 0, 1, \dots, \lfloor (M-1)/2 \rfloor\} \quad (2.15a)$$

$$d[n] := \text{fra}(\tau_p[n]) \in [-1/2, 1/2) \quad (2.15b)$$

Opisane tu postępowanie z fazą $\varphi_p[n]$ można traktować jako jej kwantowanie na M poziomów, wykorzystywane w kwantyzatorze fazy (PQ na rys.1). Zatem składniki rozkładu (2.15) można traktować jako fazę skwantowaną (2.15a) i błąd jej kwantowania (2.15b).

Ale istnieje i inna, wykorzystana dalej, interpretacja liczb występujących we wzorach (2.14) i (2.15). Zapiszmy ciąg próbek (2.10), które ma wygenerować DDS, jako

$$\exp(j\omega_M \tau_p[n]) = \exp(j\omega_M (m_p[n] + d[n])) \quad (2.16)$$

Z (2.16) widać, że $m_p[n]$ to numer próbki ciągu (2.11), (2.12) w LUT, najbliższej położeniu $\tau_p[n]$, a $d[n]$ to jej opóźnienie ułamkowe (ponieważ $|d[n]| \leq 1/2 \forall n$) względem $m_p[n]$. Do obliczania takich próbek CPS dysponuje narzędziem specjalnym: liniowym filtrem cyfrowym, zwanym filtrem ułankowo-opóźniającym (FDF-ang. *fractional-delay filter*) [6], [7]. Wykorzystamy go w bloku CRA (ang. *complex resampling algorithm*).

Jednym z ważnych zadań współczesnego CPS jest przepróbkowanie (ang. *resampling*), które polega na obliczaniu wartości próbek, które mogłyby wystąpić w chwilach innych niż oryginalne chwile próbkowania. Takie przepróbkowanie zastosowaliśmy już do konwersji szybkości próbkowania (z 44100 Sa/s na 48000 Sa/s i odwrotnie) [8], [9]. Specyfika przepróbkowania ciągu (2.11), (2.12) by otrzymać próbkę (2.16) polega na tym, że ciąg wejściowy algorytmu przepróbkowania w QO: $\exp(j\omega_M m)$, jest zespolony i skrajnie wąskopasmowy, o widmie $2\pi\delta(\omega - \omega_M)$. A zatem do jego przepróbkowywania idealnie nadaje się zespolony FDF (ang. *complex FDF* – CFDF) zaproponowany w [4]. Jest to filtr FIR o długości $N = 3, 5, \dots$, o współczynnikach zespolonych obliczanych ze wzoru:

$$\mathbf{h}_d \equiv h_d[m] = \prod_{\substack{k=-(N-1)/2 \\ k \neq m}}^{(N-1)/2} \frac{d-k}{m-k} \exp(j\omega_M (m+d)); \quad (2.17)$$

$$m = -(N-1)/2, \dots, -1, 0, 1, \dots, (N-1)/2$$

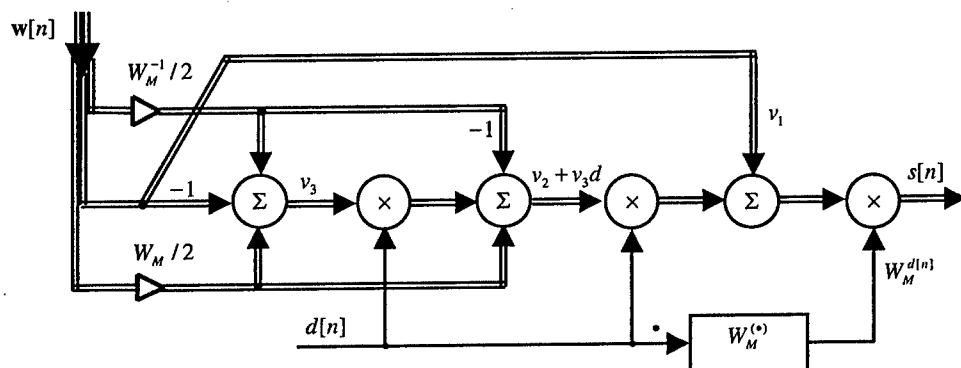
Wzór ten jest tu prezentowany w wersji nieprzyczynowej. Umożliwia to łatwą interpretację pobierania próbek z LUT (opóźnienie transportowe przyczynowego filtra FIR komplikuje tę sprawę). Przykładowo, dla $N=3$ otrzymujemy:

$$\begin{aligned} h_d[-1] &= d(d-1)(\exp(-j\omega_M(1-d)))/2, \\ h_d[0] &= (1-d^2)\exp(j\omega_M d), \\ h_d[1] &= d(d+1)(\exp(j\omega_M(1+d)))/2. \end{aligned} \quad (2.17a)$$

Długość filtra może być bardzo mała, jeżeli tylko wymiar M tablicy LUT jest dostatecznie duży. Do ilustracji wybraliśmy dość typowe $M=64$ i zupełnie wystarczające $N=3$. W bloku CRA próbkę wyjściową DDSa, oznaczoną na rys.1 przez $s[n]$, obliczamy jako:

$$\begin{aligned} s[n] &= \mathbf{h}_d[n] \mathbf{w}^T[n] \equiv h_d[n][-1] \exp(j\omega_M (m_p[n] + 1)) + \\ &+ h_d[n][0] \exp(j\omega_M m_p[n]) + h_d[n][1] \exp(j\omega_M (m_p[n] - 1)) \end{aligned} \quad (2.18)$$

Odpowiada to realizacji CFDF w strukturze transwersalnej [6] dla „czasu” m . W (2.18) wykładnik T oznacza transpozycję wektora $w[n]$ z rys.1, zawierającego trzy sąsiednie próbki z LUT: $w[n] \equiv [W_M^{m_p[n]+1}, W_M^{m_p[n]}, W_M^{m_p[n]-1}]$.



Rys.2. Struktura *flash Farrow* zespolonego filtra ułamkowo-opóźniającego (blok CRA z rys.1)

CRA przeliczający $w[n]$ na $s[n]$, o odpowiedzi impulsowej zależnej od $d[n]$, można zrealizować znacznie efektywniej w strukturze *flash Farrow* pokazanej na rys.2. Wykorzystujemy tu skrótowe oznaczenia: $W_M := \exp(j\omega_M)$ i $W_M^{d[n]} := \exp(j\omega_M d[n])$. Struktura *flash Farrow* działa natychmiastowo, bez stanów przejściowych w prawdziwym czasie n dyktowanym przez zegar DDSa.

Zagadnieniem godnym uwagi jest tworzenie LUT. Byłoby najlepiej, gdyby LUT zawierała bezbłędne wartości zespolonej sinusoidy bazowej $W_M^m = \exp(j\omega_M m/M)$. Jednakże dokładność MATLABa jest pod tym względem ograniczona. Wiedząc z doświadczenia, że błąd funkcji MATLABa $\exp(j\phi)$ rośnie za wzrostem $|\phi|$, tworzymy tablicę z 1/8 okresu tej funkcji wykorzystując symetrię Hermite'a funkcji $\exp(j\phi)$. Okazuje się, że utworzona w ten sposób LUT jest lepsza od utworzonej na podstawie tylko jednej próbki: $W_M = \exp(j\omega_M)$, przez domnażanie (w MATLABie): $W_M^{m\pm 1} = W_M^{\pm 1} W_M^m$.

3. DDS JAKO OSCYLATOR KWADRATUROWY I MODULATOR FM

W przypadku DDSa działającego jako QO, zadana „cyfrowa” (znormalizowana względem szybkości próbkowania) pulsacja chwilowa jest stała: $\omega[n] = \omega_c = \text{const}[n]$, przy czym $|\omega_c| \in (0, \pi)$, gdyż sinusoida kwadraturowa generowana przez ten DDS jest z reguły wykorzystywana do przesuwania widma (przemiany, heterodynowania) zarówno w górę, jak i w dół. Jakość tej wersji DDSa, jak i każdego innego oscylatora cyfrowego, określa się powszechnie w dziedzinie częstotliwości w kategoriach czystości widmowej generowanej przezeń sinusoidy. Miara tej czystości jest widmowy zakres dynamiczny wolny od prążków obcych (ang. *Spurious Free Dynamic Range* - SFDR), zdefiniowany jako różnica poziomów widmowych wygenerowanej sinusoidy i najwyższego prążka obcego (ang. *spurious peak* – najwyższy prążek obcy tj. o pulsacji różnej od pulsacji tej sinusoidy):

$$\text{SFDR} [\text{dBc}] = S(\omega_c) [\text{dB}] - S_{\max}(\omega)_{\omega \neq \omega_c} [\text{dB}] \quad (2.19)$$

We wzorze tym $S(\omega)$ jest widmem amplitudowym, uzyskanym na podstawie dostatecznie długiej, L -próbkowej obserwacji ciągu $s[n]$ z wyjścia DDS. Jakość tej estymaty jest zadowalająca tylko dla L pulsacji ω_c współmiernych z π , znanych z definicji L -punktowej DFT

$$\omega_k = 2\pi k / L, \quad k = \{-\lfloor L/2 \rfloor, \dots, \lfloor (L-1)/2 \rfloor\} \quad (2.20)$$

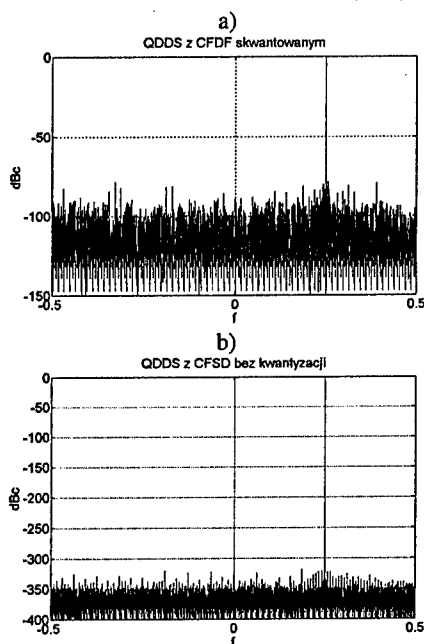
W przeciwnym przypadku zjawisko, zwane przeciekiem widma, skutecznie maskuje mniejsze szczegóły widma badanej sinusoidy. Dodajmy, że klasyczny sposób przeciwdziałania przeciekom – okienkowanie, skutkuje tu umiarkowanie i tylko przy bardzo długich obserwacjach, ujawniając wśród prążków obcych jedynie te, które są odległe od $\omega_c = 2\pi f_c$, a maskując prążki obce bliskie ω_c . Rys.3a przedstawia widmo w dBc unormowane względem wysokości najwyższego prążka DFT zespolonej sinusoidy o częstotliwości $f_c = 1023/4096$ należącej do zbioru (2.20), gdzie $\omega_c = 2\pi f_c$. Widmo obliczano dla $L=4096$ próbek DFT z wyjścia QO z kwantyzacją. Dokładność kwantyzacji w bitach (Przykład A): bits=10 dla $s[n]$, bitv1=10 dla $v_1[n]$, bitv2=5 dla $v_2[n]$, bitv3=3 dla $v_3[n]$ i bitd=6 dla $d[n]$, gdzie $v_1[n]$, $v_2[n]$ i $v_3[n]$ to wyjścia sumatorów. Dla porównania, na rys.3b pokazujemy widmo unormowane w dBc dla tego samego QO ale bez kwantyzacji, a na rys.3c dla sinusoidy zespolonej wygenerowanej w MATLABie. Widzimy, z porównania rys.3b i c, że sinusoida kwadraturowa wygenerowana za pomocą QO bez kwantyzacji ma niższy poziom prążków obcych od wygenerowanej w MATLABie. Zauważmy też, że zakres dynamiczny wolny od prążków obcych na rys.3a: SFDR=77.56 dBc, jest większy o ponad 10 dBc w porównaniu z [10], gdzie zastosowano filtr FDF rzeczywisty z [11] i te same parametry kwantyzacji. Nasz wynik jest porównywalny z wynikiem uzyskanym w [1]. Na rys.4 pokazujemy widmo unormowane dla tego samego QO z kwantyzacją, ale tu (Przykład B): bits=15, bitv1=15, bitv2=10, bitv3=5 i bitd=7, a więc dokładność kwantyzacji jest większa niż w Przykładzie A.

Uzyskany tu wynik: SFDR=105.3 dBc, jest lepszy niż opublikowany w [12], gdzie SFDR=100 dBc, ale tam zastosowano bits=16, a tu wystarczyło bits=15 na wyjściu DDSa. Dodajmy, że tak korzystnie duża wartość SFDR, jak w Przykładzie B z zespolonym FDF, jest nieosiągalna za pomocą rzeczywistego FDF o tej samej długości $N=3$. Jest tak dlatego, że nawet bez kwantyzacji (w precyzji zmiennie-przecinkowej arytmetyki MATLABa, która jest rzędu -313 dB) SFDR jest wówczas mniejszy i wynosi: 92.7 dBc, a z kwantyzacją: 71.4 dBc.

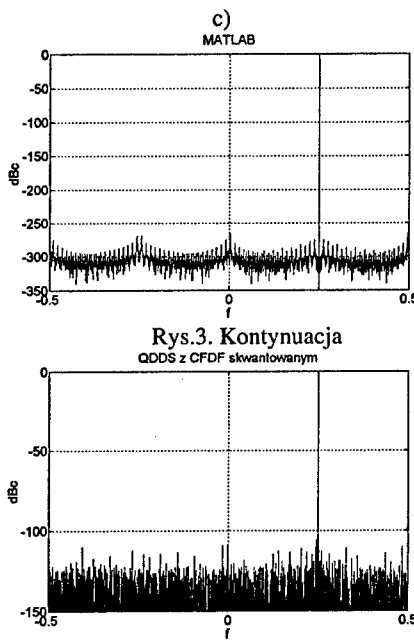
Gdy DDS pracuje jako modulator FM albo jako QO o pulsacji innej niż w (2.20), potrzebna jest inna od SFDR miara czystości wygenerowanego przebiegu $s[n]$. Brak w literaturze metody, która zastosowana do QO byłaby pozbawiona wymienionej wcześniej wady SFDR, wynikającej z przecieku widma. Wada ta nie występuje w metodzie opisanej poniżej. Bada ona czystość $s[n]$ w dziedzinie czasu, obliczając przebieg błędu chwilowego: amplitudy i pulsacji, oraz błędu zespolonego. Periodogramy tych błędów są miarą czystości $s[n]$ w dziedzinie częstotliwości. Metoda ta jest bardziej uniwersalna od SFDR. Może ona być równie miarodajnie stosowana do DDSa w roli QO, jak i modulatora FM, jak również do wzajemnego porównywania nowych opracowań DDS. Stosując tę nowo wprowadzaną miarę do QO, porównamy dalej czystość widmową przebiegów wygenerowanych przez proponowany tu DDS z przebiegami obliczonymi w MATLABie ze wzorów analitycznych (2.4) i (2.5). Do zdefiniowania wyżej wymienionych błędów chwilowych niezbędna jest

wiedza o przebiegach: amplitudy chwilowej $a[n] > 0$ i pulsacji chwilowej $\omega[n]$, które mają cechować przebieg $s[n]$ generowany przez nasz DDS (i każdy inny oscylator albo modulator FM lub AM • FM). Wcześniej posługiwaliśmy się sygnałem (2.4) o amplitudzie chwilowej równej 1. Teraz wykorzystamy ogólnie zapisany sygnał: $a[n]\exp(j\phi[n])$, gdzie $n = 0, 1, 2, \dots$. Chwilowy (bieżący) błąd amplitudy tego sygnału zdefiniujemy następująco

$$e_a[n] := |s[n]| / a[n] - 1; \quad n = 0, 1, 2, \dots \quad (2.21)$$



Rys.3. Widma unormowane na wyjściu DDS jako QO:
a) z kwantyzacją: SFDR=77.56 dBc,
b) bez kwantyzacji: SFDR=318.56 dBc
c) dla sygnału wygenerowanego bezpośrednio w MATLABie:
SFDR = 265.6 dBc



Rys.4. Widmo unormowane na wyjściu DDSa jako QO z kwantyzacją:
SFDR = 105.3 dBc

W szczególności, przy pożądanej amplitudzie o wartości jednostkowej, jak w naszym przypadku, definicja (2.21) upraszcza się do postaci

$$e_a[n] := |s[n]| - 1; \quad n = 0, 1, 2, \dots \quad (2.21a)$$

Chwilowy (bieżący) błąd pulsacji zdefiniujemy jako

$$e_\omega[n] := \begin{cases} 0, & n = 0 \\ \text{Arg}(s[n]s^*[n-1]\exp(-j\omega[n])); & n = 1, 2, 3, \dots \end{cases} \quad (2.22)$$

gdzie $\text{Arg}(\cdot)$ jest „arcustangensem cztero-ćwiartkowym” aproksymowanym w MATLABie za pomocą funkcji $\text{angle}(\cdot)$. Gdy spodziewamy się bardzo małych (w stosunku do 1 rad/Sa) błędów pulsacji (jak w naszym przypadku), to definicję (2.22) możemy uprościć do postaci

$$e_\omega[n] := \begin{cases} 0, & n = 0 \\ \text{Im}(s[n]s^*[n-1]\exp(-j\omega[n])); & n = 1, 2, 3, \dots \end{cases} \quad (2.22a)$$

(Wyodrębnienie $e_\omega[0]$ w (2.22) i (2.22a) wiąże się z przyrostową (inaczej różnicową) definicją pulsacji chwilowej: $\omega[n] := \varphi[n] - \varphi[n-1]$, gdzie $\varphi[n]$ jest fazą chwilową rozwiniętej próbki $s[n]$.) Chwilowe błędy amplitudy i pulsacji wygenerowanego przebiegu $s[n]$ można połączyć w jego jeden chwilowy błąd zespolony

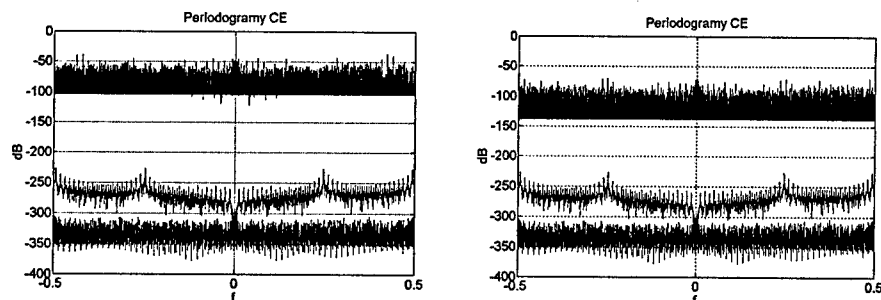
$$e_c[n] := e_a[n] + je_\omega[n] = \begin{cases} |s[n]| - 1, & n = 0 \\ |s[n]| - 1 + j \operatorname{Im}(s[n]s^*[n-1]\exp(-j\omega[n])); & n = 1, 2, 3, \dots \end{cases} \quad (2.23)$$

Błąd ten można nazwać krótko: CE od ang. *complex error*, jak na rys.5.

Wizualizacja i analiza przebiegów błędów zdefiniowanych tu w dziedzinie czasu dostarcza istotnych informacji o wygenerowanym sygnale $s[n]$, a przede wszystkim o jego stacjonarności. Natomiast wykrywanie i szacowanie wielkości prążków obcych wykonuje się łatwiej w dziedzinie częstotliwości, na podstawie widma obliczonego dla dostatecznie długiej (L -próbkowej) obserwacji przebiegu błędu zespolonego (2.23) lub jego składowych (2.21a) i (2.22a). Przykłady zastosowania opisanych tu miar jakości DDSa (kryteriów szacowania czystości przebiegu $s[n]$ wygenerowanego przez DDS) prezentujemy na rys.5. Pokazano tu periodogramy CE dla liczącej $L=4096$ próbek obserwacji sygnału $s[n]$ z QO dla Przykładu A i B (DDS o widmach z rys.3 i 4), obliczone za pomocą wzoru

$$10\log_{10} \left(F_L \{CE[n]\}^2 / L \right) [\text{dB}] \quad (2.24)$$

gdzie F_L jest operatorem L -punktowej DFT.



Rys.5. Periodogramy CE dla DDS jako QO z kwantyzacją (u góry), bez kwantyzacji (u dołu) i sygnału wygenerowanego za pomocą wzorów (2.4) i (2.5) w MATLABie (pośrodku):

- dla kwantyzacji jak w Przykładzie A (por. rys.3)
- dla kwantyzacji jak w Przykładzie B (por. rys.4).

Na rys.5, porównując dwa najniższe periodogramy, które powtarzają się na rys.5a i b widzimy, że DDS bez kwantyzacji generuje sinusoidę kwadraturową obciążoną mniejszym błędem niż ta sama sinusoida wygenerowana za pomocą wzorów (2.4) i (2.5) w MATLABie. Wysokość maksymalnego prążka tego periodogramu równa jest prawie dokładnie wartości SFDR- $10\log(L)$ jeżeli tylko liczba bitów kwantyzacji jest nie mniejsza od 8 (wtedy zanika autokorelacja szumu kwantyzacji). Analogiczne wyniki otrzymujemy dla sygnału $s[n]$ z modulacją FM. Potwierdza to użyteczność praktyczną wprowadzonej miary.

4. WNIOSKI

Zaproponowaliśmy nowy algorytm kwadraturowego DDS. Pozwala on osiągnąć zarówno wysoki stopień czystości generowanej sinusoidy kwadraturowej, jak i sygnału z modulacją FM. Wyniki eksperymentów pokazały, że dzięki zastosowaniu zespolonego, zamiast rze-

czywistego, cyfrowego filtru ułamkowo opóźniającego, przy generacji sinusoidy zespolonej można tu zwiększyć zakres dynamiczny wolny od prążków obcych nawet o 30 dBc. Za pomocą powyższego DDS uzyskujemy również bardzo małe błędy modulacji częstotliwości (FM). To wszystko osiągamy przy małej pojemności pamięci ROM, od której zależy pobór mocy zasilania. Ponadto zaproponowaliśmy i zilustrowaliśmy nowe podejście do ilościowej oceny jakości DDS, odpowiednie nie tylko do oceny czystości generowanej zespolonej sinusoidy, ale również błędów modulacji.

BIBLIOGRAFIA

- [1] Bellaouar A., O'brecht M.S., Fahim A.M. and Elmasry M.I.: *Low-power DDFS for wireless communications*. IEEE Journal of Solid-State Circuits, vol. 35, No.3, 2000, pp. 385-390.
- [2] Sodgar A.M and Lahiji G.R.: *A pipelined ROM-less architecture for sine output DDFSs using the second order parabolic approximation*. IEEE Trans. on Circuits and Systems-II: Analog and Digital Signal Processing, vol. 48, No. 9, 2001, pp. 850-857.
- [3] Langlois J.M.P. and Al-Khalili D., *Novel approach to the design of DDFSs based on linear interpolation*, IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, vol. 50, No. 9, 2003, pp. 567-578.
- [4] Hermanowicz E.: *Explicit formulas for weighting coefficients of maximally flat tunable FIR delayers*. Electronics Letters, vol. 28, No. 20, 1992, pp. 1936-1937.
- [5] Langlois J.M.P. and Al-Khalili D., *A low power direct DDFS with 60dBc spectral purity*. The 12th ACM Great Lakes Symposium on VLSI, New York, April 18-20, 2002, pp. 166-172.
- [6] Laakso T., Valimäki V., Karjalainen M. and Laine U.K.: *Splitting the unit delay. Tools for fractional delay filter design*. IEEE Signal Processing Magazine, January 1996, pp. 30-60.
- [7] Hermanowicz E.: *Specjalne filtry dyskretne o skończonej odpowiedzi impulsowej i ich zastosowania do modulacji i demodulacji kwadraturowej*. ELEKTRONIKA, No 82, Gdańsk 1995.
- [8] Hermanowicz E., Rojewski M. and Blok M.: *A sample rate converter based on a fractional delay filter bank*. W: Proceedings of International Conference on Signal Processing Applications and Technology, ICSPAT 2000, Dallas, Texas, USA, October 16-19, 2000; publikacja elektroniczna.
- [9] Blok M.: *Projektowanie opóźniających filtrów cyfrowych FIR metodą iteracji czasowo-częstotliwościowej*, Rozprawa doktorska, WETI Politechnika Gdańska, Gdańsk 2003.
- [10] Eltavil A.A. and Daneshrad B.: *Piece-wise parabolic interpolation for DDFS*. W: Proceedings of the IEEE Custom Integrated Circuits Conference, 2002, pp. 401-404.
- [11] Erup L., Gardner R.M. and Harris R.A.: *Interpolation in digital modems – Part II: Implementation and performance*. IEEE Transactions on Comm., vol. 41, 1993, pp. 998-1008.
- [12] Madiseti A., Kwentus A. and Willson A.N.: *A 100-MHz, 16-b, DDFS with a 100-dBc spurious-free dynamic range*. IEEE Journal on Solid-State Circuits, vol. 34, No. 8, 1999, pp. 1034-1043.

QUADRATURE DDS WITH FLASH-FARROW FRACTIONAL-DELAY FILTER

Summary

The subject of the paper is a nonlinear discrete-time algorithm of direct digital synthesizer (DDS). We propose a new quadrature DDS algorithm using very small ROM, which leads to small power consumption. It allows to reach high purity of generated complex sinusoid as well as very small FM-modulation errors. Furthermore we propose a new criterion for quantitative evaluation of the DDS quality. The criterion is adequate equally well to the assessment of the generated complex sinusoid and the FM-modulation errors.

Michał Jacymirski*, Piotr Lipiński**

***Instytut Informatyki, Politechnika Łódzka**

****Katedra Mikroelektroniki i Technik Informatycznych, Politechnika Łódzka**

EFEKTYWNY ADAPTACYJNY ALGORYTM TRANSFORMATY FALKOWEJ DAUBECHIES 4

Streszczenie

W artykule zaproponowano nowe podejście do obliczania adaptacyjnych pakietów falkowych umożliwiające dopasowanie funkcji bazowej do rodzaju przetwarzanego sygnału na każdym etapie transformaty. Przedstawiono efektywny algorytm obliczania takiej transformaty na przykładzie transformaty Haara i transformaty Daubechies 4. Wykazano, że złożoność obliczeniowa przedstawionego algorytmu jest mniejsza od złożoności obliczeniowej tradycyjnego algorytmu obliczania transformaty falkowej Daubechies 4 przy pomocy splotu.

1. WSTĘP

Adaptacyjne pakiety falkowe są bardzo ważną odmianą dyskretnych transformat falkowej, w której zachowuje się wyniki transformaty zawierające najważniejsze informacje z punktu widzenia użytkownika. Podczas obliczania adaptacyjnych pakietów falkowych dokonuje się oceny zawartości informacji w sygnale po każdym etapie transformaty jednego rodzaju. Do dalszych obliczeń wybierane są tylko te ciągi, które spełniają założone kryteria.

Przy opracowywaniu transformat falkowych dąży się do wybrania takiej funkcji bazowej, aby miała ona charakter zbliżony do charakteru analizowanego sygnału [2], [6], [8], [11]. W zastosowaniach praktycznych stosuje się eksperymentalny dobór falki na podstawie wyników transformaty [2], [13]. W przypadku adaptacyjnych pakietów falkowych, aby wybrać bank filtrów najlepiej dopasowany do właściwości przetwarzanego sygnału, należy kilkakrotnie wykonać procedurę adaptacyjną, co wymaga kilkakrotnego przekształcania tych samych danych i nie jest efektywne.

Dla usunięcia tej wady w artykule zaproponowano nowe podejście do konstruowania adaptacyjnych pakietów falkowych, w którym przekształcanie różnych rodzajów wykonuje się jednocześnie. Przy czym, przekształcenie jednego rodzaju realizowane jest na podstawie wyników przekształcenia innego rodzaju, a nie na podstawie sygnału wejściowego. Jest to możliwe dzięki zastosowaniu nowego podejścia do obliczania transformaty falkowej [7], [14], w którym odpowiedź impulsowa filtrów podana jest jako suma trywialnych

ciągów. Złożoność prezentowanego algorytmu jest mniejsza od złożoności obliczeniowej tradycyjnego algorytmu obliczania transformaty falkowej Daubechies 4 [10] i jest równa szybkiemu algorytmowi przedstawionemu w [7]. Istotne jest, że dzięki zastosowaniu nowatorskiego podejścia do obliczania transformaty falkowej Daubechies 4, obliczenie transformaty falkowej Haara nie wymaga zwiększenia liczby operacji matematycznych.

2. TRANSFORMATA FALKOWA HAARA I TRANSFORMATA FALKOWA DAUBECHIES 4

Transformata falkowa Haara jest historycznie pierwszą i jednocześnie najprostszą transformatą falkową. Znajduje ona zastosowanie w kompresji, poprawie kontrastu i klasyfikacji obrazów medycznych [3],[5],[13]. Dyskretną transformatę Haara oblicza się przy użyciu banku filtrów (2.1), (2.2). Dokładny opis właściwości transformaty Haara oraz algorytmów obliczeniowych został zawarty w [1].

$$D = [1, 1] \quad (2.1)$$

$$G = [1, -1] \quad (2.2)$$

gdzie:

D – odpowiedź impulsowa filtru dolnoprzepustowego,

G – odpowiedź impulsowa filtru górnoprzepustowego.

Transformata falkowa Daubechies jest szeroko stosowana. Posiada ona odmianę zarówno dyskretną jak i ciągłą, co znacznie ułatwia interpretację wyników obliczeń. Jest transformatą ortogonalną i może być obliczana z zastosowaniem banku filtrów [1]. Wśród transformat falkowych Daubechies często stosuje się transformatę Daubechies 4. Liczba 4 oznacza liczbę współczynników odpowiedzi impulsowej filtrów. Współczynniki odpowiedzi impulsowej filtrów transformaty falkowej Daubechies 4 mogą zostać obliczone kilkoma sposobami [4],[12],[15]. W rozpatrywanym przypadku można zastosować dowolną metodę, konieczne jest jednak obliczenie współczynników w postaci analitycznej. Odpowiedzi impulsowe filtrów Daubechies 4 w postaci analitycznej wynoszą:

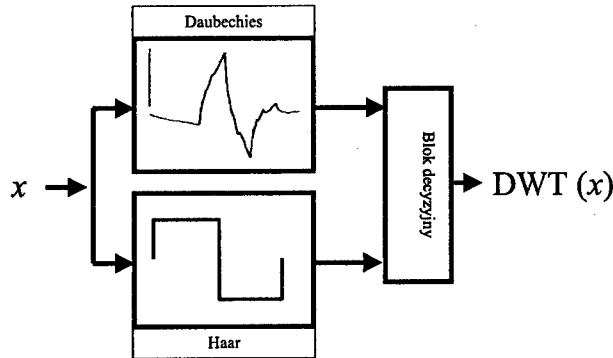
$$D = \left[\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}} \right] \quad (2.3)$$

$$G = \left[\frac{1-\sqrt{3}}{4\sqrt{2}}, \frac{-3+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{-1-\sqrt{3}}{4\sqrt{2}} \right] \quad (2.4)$$

3. EFEKTYWNE ADAPTACYJNE PAKIETY FALKOWE Z AUTOMATYCZNYM WYBOREM FUNKCJI BAZOWEJ

W algorytmie adaptacyjnych pakietów falkowych dokonuje się oceny sygnału po każdym etapie obliczeniowym. Jeżeli założone kryterium jest spełnione, obliczanie transformaty jest kontynuowane, jeżeli nie, dalsze etapy transformaty nie są obliczane. Ocena sygnału dokonywana w algorytmie adaptacyjnych pakietów falkowych może dodatkowo posłużyć do wyboru funkcji bazowej, która najlepiej odpowiada przetwarzanemu sygna-

łowi. Na obecnym etapie badań, wyboru dokonuje się między dwoma funkcjami bazowymi Haara i Daubechies 4, co schematycznie zostało przedstawione na rys 1.



Rys. 1. Schemat ideowy algorytmu obliczania adaptacyjnych pakietów falkowych z automatycznym wyborem funkcji bazowej

Do oceny wyników transformaty Haara i transformaty Daubechies, konieczne jest obliczenie obu transformat a następnie porównanie obu ciągów wynikowych w oparciu o wybrane kryterium. Podejście takie nie jest jednak efektywne pod względem obliczeniowym. Wymaga bowiem obliczenia obydwu transformat, w celu ich porównania. Dublowania obliczeń można uniknąć korzystając z własności filtrów transformaty falkowej Daubechies opisanych w pracy [7]. Aby wykazać tę własność należy przekształcić filtry transformaty falkowej (2.3) i (2.4) do postaci danej zależnościami odpowiednio (3.1), (3.2):

$$D = \frac{1}{4\sqrt{2}} \left([-1, 1, 1, -1] + 2[1, 1, 1, 1] + \sqrt{3}[1, 1, -1, -1] \right) \quad (3.1)$$

$$G = \frac{1}{4\sqrt{2}} \left([-1, -1, 1, 1] + 2[1, -1, 1, -1] + \sqrt{3}[-1, 1, 1, -1] \right) \quad (3.2)$$

Stała $\frac{1}{4\sqrt{2}}$ może zostać pominięta w procesie obliczeniowym, ponieważ powoduje ona jedynie przeskalowanie sygnału wyjściowego. Obliczenie transformaty falkowej Daubechies 4 sprowadza się zatem do obliczenia wyrażeń znajdujących się w nawiasach zależności (3.1) i (3.2). Operacja mnożenia skalarnego ciągu $x(n)$ przez filtr (wektor) o odpowiedzi impulsowej przyjmującej wartości 1 i -1 może być traktowana jako suma lub różnica kolejnych elementów ciągu, co prowadzi do zależności (3.3), (3.4):

$$\begin{cases} d_{Di}(n) = (-x(2n) + x(2n+1) + x(2n+2) - x(2n+3)) + \\ \quad + 2(x(2n) + x(2n+1) + x(2n+2) + x(2n+3)) + \sqrt{3}(x(2n) + x(2n+1) - x(2n+2) - x(2n+3)), \end{cases} \quad (3.3)$$

$$\begin{cases} g_{Di}(n) = (-x(2n) - x(2n+1) + x(2n+2) + x(2n+3)) + \\ \quad + 2(x(2n) - x(2n+1) + x(2n+2) - x(2n+3)) + \sqrt{3}(-x(2n) + x(2n+1) + x(2n+2) + x(2n+3)), \end{cases} \quad (3.4)$$

Dokonując podstawień $a(k) = (x(2n) + x(2n+1))$ oraz $b(k) = (x(2n) - x(2n+1))$ równania (3.3), (3.4) mogą zostać przekształcone do postaci (3.5), (3.6):

$$\begin{cases} d_i(n) = (-b(n) + b(n+1)) + 2(a(n) + a(n+1)) - \sqrt{3}(a(n) - a(n+1)), & (3.5) \\ g_i(n) = (-a(n) - a(n+1)) + 2(b(n) + b(n+1)) + \sqrt{3}(-b(n) + a(n+1)) & (3.6) \end{cases}$$

Należy podkreślić, że obliczenie zależności $a(n)$ oraz $b(n)$ jest tożsame z obliczeniem transformaty Haara (2.1), (2.2). Z zależności (3.5), (3.6) wynika, że transformaty falkowej Daubechies 4 można dokonać wykorzystując wyniki uprzednio obliczonej transformaty Haara. Fakt ten leży u podstaw opracowania efektywnego algorytmu obliczania adaptacyjnych pakietów falkowych. Liczba operacji matematycznych niezbędnych do obliczenia algorytmu może zostać dodatkowo zmniejszona poprzez zastosowanie podstawień: $A(n) = (a(n) - a(n+1))$ i $B(n) = (b(n+1) - b(n))$, co prowadzi do zależności (3.7), (3.8):

$$\begin{cases} d_i(n) = (B(n)) + 2(a(n) + a(n+1)) - \sqrt{3}(A(n)), & (3.7) \\ g_i(n) = (A(n)) + 2(b(n) + b(n+1)) + \sqrt{3}(B(n)) & (3.8) \end{cases}$$

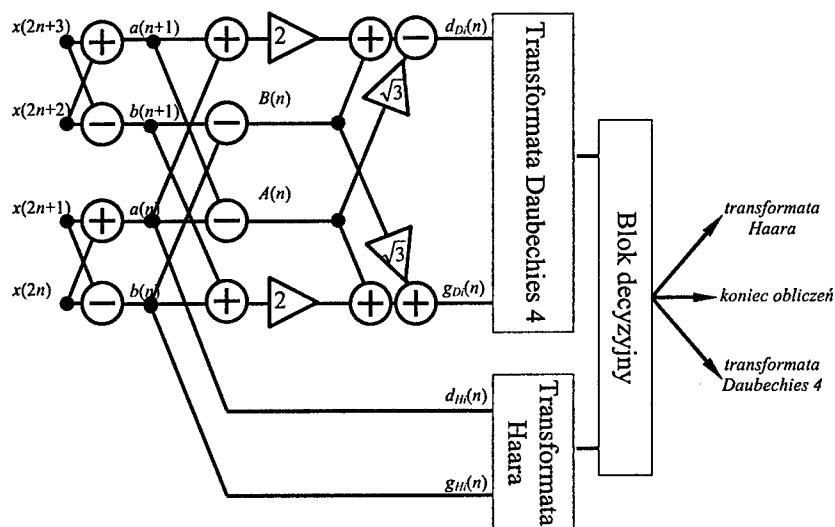
Ostatecznie, efektywny algorytm można podzielić na następujące etapy:

- obliczenie transformaty Haara sygnału wejściowego ($a(n)$ i $b(n)$),
- obliczenie $A(n)$ i $B(n)$,
- obliczenie transformaty Daubechies 4 na podstawie wyników transformaty Haara (zależności (3.7), (3.8)),
- porównanie wyników transformat, wybór lepszego wariantu w oparciu o wybrane kryterium,
- podjęcie decyzji o kontynuacji obliczania transformaty.

Schemat obliczeniowy algorytmu adaptacyjnego z wyborem funkcji bazowej został przedstawiony na rys. 2. Schemat ilustruje zależność między transformatą Haara i transformatą Daubechies. Jak łatwo zauważyć, transformata Haara jest obliczana jako jedna z sum pośrednich w algorytmie obliczeniowym transformaty Daubechies 4. W bloku decyzyjnym dokonywany jest wybór spośród trzech możliwości:

- wybór transformaty Daubechies 4 do dalszych obliczeń,
- wybór transformaty Haara do dalszych obliczeń,
- koniec obliczeń.

Kryterium wyboru zależy od charakteru przetwarzanego sygnału oraz jego zastosowania. Powszechnie stosowanym kryterium jest entropia sygnału, która może być również stosowana w tym przypadku [9].



Rys. 2. Schemat obliczeniowy efektywnego algorytmu obliczania adaptacyjnych pakietów falkowych z automatycznym wyborem funkcji bazowej

4. OCENA ZŁOŻONOŚCI OBLICZENIOWEJ ALGORYTMU

Złożoność obliczeniowa zaproponowanego algorytmu została oszacowana bez uwzględnienia obliczeń przeprowadzanych w bloku decyzyjnym. Jest to uzasadnione, ponieważ liczba operacji matematycznych konieczna do podjęcia decyzji nie zależy od zastosowanego algorytmu obliczania transformaty a jedynie od przyjętego kryterium. Liczba wykonanych operacji matematycznych związanych z blokiem decyzyjnym jest zatem identyczna niezależnie od przyjętej metody obliczania transformaty. Należy podkreślić, że w tradycyjnym algorytmie pakietów falkowych, w którym nie ma możliwości wyboru różnych funkcji bazowych, zastosowanie bloku decyzyjnego jest konieczne by stwierdzić, czy obliczenia mają być kontynuowane czy też zakończone. Złożoność obliczeniowa bloku decyzyjnego tradycyjnej metody pakietów falkowych jest tylko nieznacznie mniejsza w stosunku do tego zaproponowanego w algorytmie.

Oszacowania złożoności numerycznej dokonano dla jednego etapu transformaty. Jako jeden etap rozumie się obliczenie transformat Haara (d_{Hi} , g_{Hi}) i Daubechies 4 (d_{Di} , g_{Di}) na podstawie ciągu wejściowego x . Przyjęto, że ciąg wejściowy składa się z l_x elementów. Pierwszym etapem algorytmu jest obliczenie transformaty Haara sygnału x , co wymaga l_x dodawań/odejmowań. Do obliczenia transformaty Daubechies 4 wymagane jest dodatkowo $6l_x$ dodawań i $2l_x$ mnożeń. W bezpośrednim algorytmie transformat falkowych obliczenie jednego etapu transformaty wymaga $8l_x$ mnożeń i $6l_x$ dodawań [7], [10]. Zestawienie liczby operacji niezbędnych do obliczenia jednego etapu transformaty zostało przedstawione w tabeli 1.

Tabela 1

Porównanie złożoności obliczeniowej algorytmów

	algorytm bezpośredni	algorytm efektywny	Zysk
Liczba dodawań/odejmowań	$6l_x$	$8l_x$	$-2l_x$
liczba mnożeń	$8l_x$	$2l_x$	$6l_x$

Przewaga przedstawionego algorytmu w stosunku do algorytmu tradycyjnego jest szczególnie widoczna, gdy operacji obliczania transformaty falkowej dokonuje się przy użyciu procesora wyposażonego w pojedynczą jednostkę arytmetyczno logiczną. Jest to spowodowane następującymi czynnikami:

- przewagą liczby mnożeń nad liczbą dodawań w algorytmie tradycyjnym,
- większą złożonością obliczeniową mnożenia w stosunku do dodawania i odejmowania (w układzie wyposażonym w pojedynczą jednostkę logiczną, mnożenie liczb zmiennoprzecinkowych, k -bitowych wymaga ok. k razy więcej operacji niż wykonanie dodawania lub odejmowania),
- czterokrotną redukcją liczby mnożeń w przedstawionym algorytmie.

5. WNIOSKI

W artykule przedstawiono adaptacyjny, efektywny algorytm obliczania pakietów falkowych z możliwością automatycznego dopasowania funkcji bazowej. Wykazano związek między transformatą falkową Haara oraz transformatą falkową Daubechies 4. Wykorzystując ten związek zaproponowano metodę obliczania transformaty falkowej Daubechies 4, która oparta jest na wynikach przekształcenia Haara. Zaproponowana metoda posłużyła jako podstawa do opracowania adaptacyjnego algorytmu obliczania pakietów falkowych z automatycznym wyborem funkcji bazowej między funkcją Haara i Daubechies 4. Ponadto wykazano, że złożoność numeryczna opracowanego algorytmu jest mniejsza niż złożoność numeryczna algorytmu bezpośredniego, który nie daje możliwości wyboru funkcji bazowej. Złożoność obliczeniowa otrzymanego algorytmu jest równa złożoności obliczeniowej algorytmu efektywnego Daubechies 4 przedstawionego w [7].

W dalszych pracach, autorzy mają zamiar rozwinąć przedstawioną metodę tak, aby umożliwić wybór funkcji bazowych spośród funkcji Daubechies wyższych rzędów, zachowując jednocześnie małą złożoność obliczeniową algorytmu.

BIBLIOGRAFIA

- [1] Białasiewicz J.T.: *Falki i Aproksymacje*, WNT Warszawa 2000 ss.250
- [2] Chapa J.O., Rao R.M.: *Algorithms for Designing Wavelets to Match a specified Signal*. W: IEEE Trans. on Sig. Proc., Vol. 48, No. 12, 2000, s. 3395-3406.
- [3] Coifman R. Wickerhauser M. V.: *Experiments with Adapted Wavelet De-noising for Medical Signals and Images/ Time Frequency and Wavelets in Biomedical Signal Processing*, New York IEEE Press, 1997.-s. 323-365.
- [4] Daubechies I.: *Ten lectures on Wavelets*. W: CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pensylvania, 1992.

- [5] Fusheng Y., Bo H. Qingyu T.: *Approximate Entropy and Its Application in Biosignal Analysis/Nonlinear Biomedical Signal Processing Dynamic Analysis and Modeling* Vol. II, IEEE, New York IEEE Press, 2001, s. 72-92.
- [6] Ho K.C., Chan Y.T.: *An Iterative Algorithm for Two Scale Wavelet Decomposition*. W: IEEE Trans. on Sig. Proc., Vol. 49, No. 1, 2001, s. 254-257.
- [7] Lipiński P.: *Fast Algorithm For Daubechies Discrete Wavelet Transform Computation*. W: Modelowanie i Technologie Informacyjne, Zbiór prac naukowych no. 19, Kijów 2002, s. 178-183.
- [8] Mallat S.: *A Wavelet Tour of Signal Processing*, Academic Press, 1998, pp
- [9] Misiti M., Oppenheim G., Poggi J.-M., Matlab 6 Release 12, The MathWorks Inc. 18 maj 2001, Wavelet Toolbox Help
- [10] Numerical Recipes in C: The art of Scientific Computing, 1988-1992 by Cambridge University Press, (ISBN 0-521-43108-5), pp 591-606.
- [11] Romaniuk P.: *Zastosowanie filtracji nieliniowej w dziedzinie transformaty falkowej do redukcji zakłóceń mięśniowych w sygnale elektrokardiograficznym*. Praca Doktorska, Politechnika Łódzka, Wydział Elektrotechniki i Elektroniki, Instytut Elektroniki, Łódź 2002.
- [12] Skarbek W.: *Multimedia Algorytmy i standardy kompresji*. Akademicka Oficyna Wydawnicza PLJ, Warszawa, 1998.
- [13] Walter B. Richardson Jr.: *Wavelets Applied to Mammograms/Time Frequency and Wavelets in Biomedical Signal Processing*. New York IEEE Press, 1997.-s. 499-518.
- [14] Yatsymirskyy M.: *The Fast Orthogonal Trigonometric Transform Algorithms*. Lviv: Academic Express LTD, 1997.
- [15] Zieliński T.P. *Od teorii do cyfrowego przetwarzania sygnałów*, ANTYKWA Kraków 2002.

EFFECTIVE ALGORITHM FOR ADAPTIVE DAUBECHIES 4 WAVELET TRANSFORM COMPUTATION

Summary

In this paper a new approach to wavelet packets computation is introduced. A new transform including adaptive algorithm for wavelet basis function selection is proposed. In this approach a wavelet basis function can be selected at each level of wavelet transform. An effective algorithm for the transform computation is also introduced. The number of arithmetical operations of the new algorithm is compared with the number of numerical operations of traditional one, basing on convolution.

Przemysław Maziewski

Katedra Systemów Multimedialnych, Politechnika Gdańska

ALGORYTM NORMALIZACJI POZIOMÓW GŁOŚNOŚCI DZWIĘKU ZAREJESTROWANEGO W PLIKACH

Streszczenie

W pracy przedstawiono algorytm normalizacji głośności plików dźwiękowych dedykowany głośnikowemu odsłuchowi wielokanałowemu. Algorytm generuje wartości wzmacnień potrzebne do ujednolicenia głośności plików dźwiękowych. Są one uzyskiwane na podstawie normalizacji wartości skutecznej, odpowiednio przefiltrowanych plików dźwiękowych. W celu wyeliminowania ewentualnych przesterowań, po wykonanej normalizacji wartości skutecznej, następuje proces normalizacji wartości szczytowych. W ostatniej części pracy przedstawiono wyniki dwóch testów, za pomocą których zbadano algorytm. Dodatkowo przedstawiono również krótki opis zjawiska postrzegania głośności z uwzględnieniem obiektywnych i mierzalnych wielkości wykorzystywanych w modelach jego opisu.

1. WSTĘP

Proces postrzegania głośności jest zjawiskiem złożonym. Do czynników na nie wpływających należy zaliczyć: strukturę widmową, poziom ciśnienia dźwięku i jego zmiany w czasie (spowodowane modulacją amplitudy). Autorzy publikacji związanych z tą tematyką wskazują kilka mierzalnych wielkości opisujących sygnał dźwiękowy, które mogą posłużyć do opisu subiektywnie postrzeganej głośności [1, 2].

Postrzeganie głośności zależne jest od zmian sygnałów w czasie, które przy przyjęciu odpowiednich uproszczeń można nazwać modulacją amplitudy. W przypadku niewielkich częstotliwości modulacyjnych postrzegany poziom głośności może być opisany za pomocą relacji pomiędzy skutecznym poziomem sygnału a jego chwilową wartością szczytową. W przypadku nieco większych częstotliwości modulacyjnych zmiany głośności można opisać za pomocą wartości skutecznej (pomniejszonej o pewną niewielką stałą). Wysokie częstotliwości modulujące mogą spowodować wzrost subiektywnie postrzeganej głośności sygnału [1].

Kolejnym czynnikiem wpływającym na percepcję głośności jest czas trwania dźwięku. W przypadku dźwięku o stałym poziomie, postrzeganie jego głośności zależne jest od jego czasu trwania [1, 2].

Innymi czynnikami wpływającymi na opisywane zjawisko są efekty filtrujące głowy, ramion i torsu oraz małżowiny i kanału usznego [1]. Ich wpływ na postrzeganą głośność

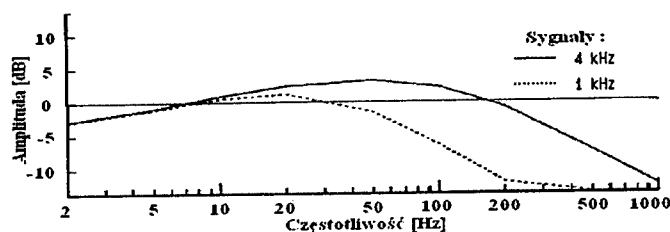
zależny jest od wzajemnego położenia źródła dźwięku i słuchacza w danej przestrzeni odsłuchowej [1, 3, 4].

Oczywistym czynnikiem wpływającym na postrzeganie głośności dźwięku są zaburzenia słuchu. W niniejszym artykule nie będą one brane pod uwagę.

Na przestrzeni ostatnich kilkunastu lat powstało kilka modeli opisujących proces postrzegania głośności. Obejmują one zarówno sygnały stałe jak i zmieniające się w czasie (modulowane amplitudowo) [1, 2, 5]. Jedną z ostatnich propozycji są dwa modele stworzone przez Glasberga, Moora i Baera [1, 2].

2. ALGORYTM NORMALIZACJI POZIOMÓW GŁOŚNOŚCI

Jako parametr opisujący głośność wybrano wartość skuteczną skończonego sygnału. Jest to świadome założenie mające na celu uprościć strukturę algorytmu. Rysunek 1 przedstawia wykresy pokazujące różnice w poziomach skutecznych, modulowanych i niemodulowanych sygnałów o jednakowej głośności (linia prosta – sygnał niemodulowany). Oś pozioma pokazuje wzrost częstotliwości modulującej, oś pionowa różnice poziomów (modulowanego i niemodulowanego sygnału).



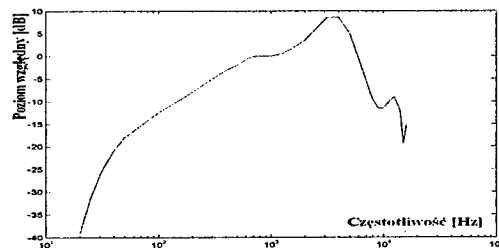
Rys.1. Różnice w poziomach skutecznych, wymagane by otrzymać jednakową głośność modulowanych i niemodulowanych sygnałów [1].

Główna idea algorytmu opiera się na próbie ujednolicenia poziomów głośności sygnałów poprzez wyrównanie ich wartości skutecznych. Określenie wartości *rms* poszczególnych sygnałów poprzedzone jest ich odpowiednią filtracją. Ma ona na celu uwzględnienie wpływu czynników takich jak: położenie źródła dźwięku w stosunku do słuchacza, efekty filtrujące ramion, torsu oraz małżowiny i kanału usznego. W filtracji wykorzystywane są odpowiedzi impulsowe, rejestrowane w komorze bezechowej za pomocą mikrofonów umieszczonych w kanałach usznych manekina. W literaturze anglojęzycznej odpowiedzi te noszą nazwę *HRIR*¹ (ang. *head related impulse responses*) [3, 4, 6].

W opisywanym algorytmie wykorzystuje się dwa zestawy odpowiedzi impulsowych. Pierwszy to standardowe funkcje *HRIR* zarejestrowane przez Keitha i Gardnera [6]. Na schemacie (rys.3.) oznaczony jest on jako „*HRIR* pole swobodne”. Drugi zestaw zawiera odpowiednio przekształcone odpowiedzi impulsowe. Takie przekształcenia polegają na normalizacji wszystkich odpowiedzi impulsowych względem jednej z nich lub względem średniej, policzonej na podstawie całego zestawu. Druga metoda określona jest jako ang. *diffuse-field equalization*. W badaniach algorytmu wykorzystano odpowiedzi znormalizowane tą właśnie metodą (stąd też nazwa „*HRIR* pole rozproszone”) [6].

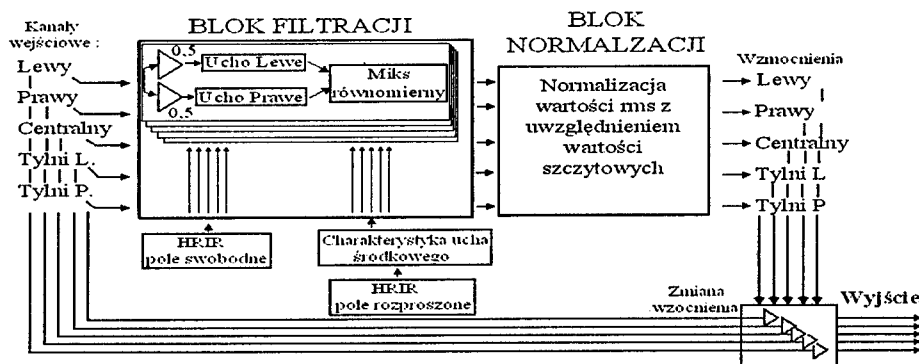
¹ Wybór odpowiednich funkcji *HRIR* podyktowany jest położeniem głośników w pomieszczeniu odsłuchowym zgodnym z normą ITU-R [7].

Proces normalizacji odpowiedzi impulsowych pozwala wyeliminować wpływ charakterystyk częstotliwościowych poszczególnych urządzeń w torze rejestracji *HRIR*. W użytych odpowiedziach jest to niepożądana różnica pomiędzy charakterystykami mikrofonu w kanałach usznych manekina i mikrofonu użytego do zbadania charakterystyki częstotliwościowej głośnika, którym odtwarzano pobudzenie [6]. Niestety proces normalizacji eliminuje również wpływ efektu filtrującego ucha zewnętrznego i środkowego. Dlatego też w przypadku stosowania zestawu „*HRIR* pole rozproszone” odpowiedzi impulsowe są dodatkowo filtrowane charakterystyką ucha zewnętrznego i środkowego. Taka charakterystyka przedstawiona jest na rysunku 2 [1].



Rys.2. Charakterystyką ucha zewnętrznego i środkowego [1].

Po filtracji następuje określenie wartości skutecznych sygnałów. W bloku normalizacji najpierw liczone są wzmacnienia potrzebne do uzyskania jednolitego poziomu *rms* wszystkich przefiltrowanych sygnałów. Następnie, by nie dopuścić do ewentualnych przesterowań, badany jest maksymalny poziom szczytowy w zbiorze znormalizowanych w ten sposób plików. Dalej wartości wzmacnień uzyskane w pierwszym kroku skalowane są współczynnikiem równym odwrotności maksymalnego poziomu szczytowego. Zmodyfikowane w ten sposób współczynniki wzmacnień pozwalają na uzyskanie jednolitych poziomów *rms* nie wprowadzając przesterowania (poziom szczytowy jest zawsze mniejszy od wartości 1). W ostatnim etapie poziomy oryginalnych (nieprzefiltrowanych) sygnałów modyfikowane są zgodnie z otrzymanymi współczynnikami wzmacnień. Rysunek 3 przedstawia blokowy schemat algorytmu.



Rys.3. Schemat blokowy algorytmu normalizacji głośności.

W „bloku filtracji” monofoniczne ścieżki dźwiękowe są równomiernie dzielone. Kolejnym krokiem jest ich filtracja w oparciu o charakterystyki częstotliwościowe otrzymane na podstawie wybranych odpowiedzi *HRIR*. Wcześniejszy podział ścieżki na ucho lewe i prawe, pozwala na symulację słyszenia dwuuszego. Całość przetwarzania w „bloku filtracji” zakończona jest równomiernym miksem sygnałów. Równy podział na ucho lewe i prawe ma swe uzasadnienie w psychoakustyce [5, 8].

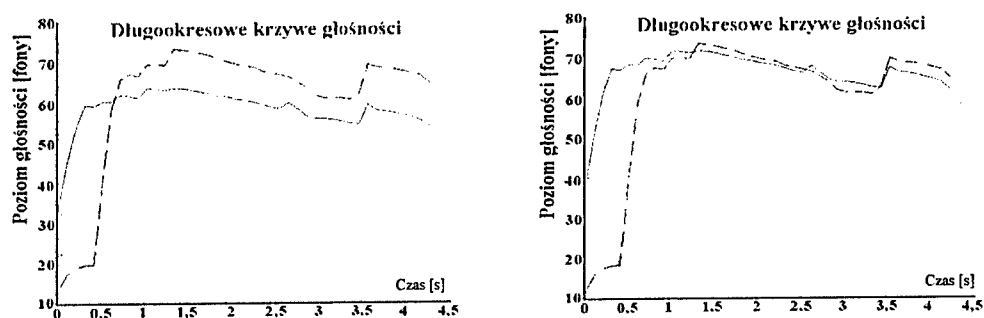
3. PRZEPROWADZONE BADANIA

W celu zbadania przydatności opisanego algorytmu poddano go dwóm testom. Pierwszy z nich wykorzystuje model głośności zaproponowany przez Glasberga i Moora [1, 2]. Dzięki temu ma on znamiona testu obiektywnego. Drugi ze wspomnianych testów opiera się na subiektywnych ocenach ekspertów.

3.1 Test wykorzystujący model głośności

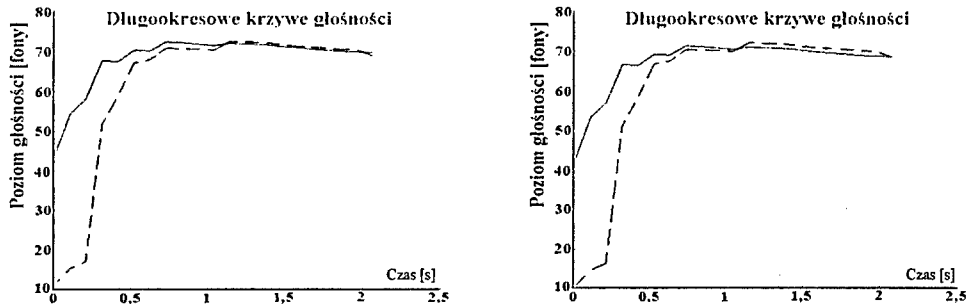
Test obiektywny wykorzystuje długookresowe krzywe głośności generowane na podstawie modelu oceny głośności zaproponowanego przez Glasberga i Moora [1, 2]. Został on zaimplementowany w środowisku *Matlab* [9].

Prezentowane wykresy pokazują krzywe głośności przed i po normalizacji wykonanej za pomocą opisywanego algorytmu. Poszczególne wykresy obrazują głośność stereofonicznego pliku dźwiękowego, w którym kanał lewy (linia ciągła) i kanał prawy (linia przerywana) zawierają tę samą wypowiedź słowną zarejestrowaną za pomocą oddzielnych urządzeń rejestrujących. Rysunek 4 przedstawia pierwszy z badanych przypadków. Wykres przedstawiający przypadek po normalizacji (prawa strona rysunku) pokazuje, iż głośność pomiędzy dwoma kanałami została w dużym stopniu ujednoliconą.



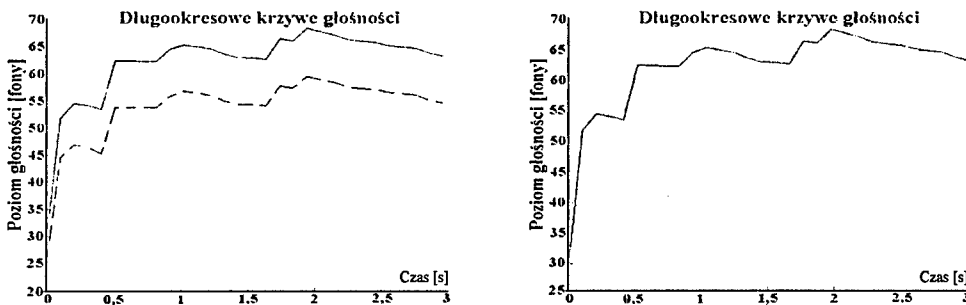
Rys.4. Wykresy obrazujące długookresowe krzywe głośności przed (lewa strona) i po normalizacji (prawa strona) wykonanej za pomocą opisywanego algorytmu.

Rysunek 5 przedstawia drugi z badanych przypadków. Jest on ciekawy z tego powodu, że w badanym sygnale wejściowym nie ma postrzegalnej różnicy w kanale lewym i prawym (lewy wykres na rysunku 5 pokazuje niewielkie różnice w głośności obu kanałów). Krzywe głośności uzyskane po normalizacji pokazują, iż relacje pomiędzy dwoma kanałami zostały zachowane. Potwierdza to poprawność działania algorytmu.



Rys.5. Wykresy obrazujące długookresowe krzywe głośności przed (lewa strona) i po normalizacji (prawa strona) wykonanej za pomocą opisywanego algorytmu.

Rysunek 6 przedstawia trzeci z badanych przypadków. W odróżnieniu od dwóch poprzednich, plik wejściowy zawiera wypowiedź słowną zarejestrowaną jednym urządzeniem. Różnice w głośności pomiędzy oba kanałami wynikają z wprowadzonego do kanału prawego tłumienia. Wykres obrazujący krzywe głośności po normalizacji pokazuje ich całkowitą zbieżność (głośność obu kanałów została ujednolicona).



Rys.6. Wykresy obrazujące długookresowe krzywe głośności przed (lewa strona) i po normalizacji (prawa strona) wykonanej za pomocą opisywanego algorytmu.

W ramach testu wykonano jeszcze kilka udanych prób normalizacji głośności analogicznych jak w trzecim z opisywanych przypadków. Różnice wynikały jedynie z wartości tłumień wprowadzanych pomiędzy kanałami. W każdym z przypadków rezultat działania algorytmu był taki sam jak w prezentowanym przykładzie (głośność obu kanałów została ujednolicona).

3.2 Test subiektywny bazujący na ocenach ekspertów

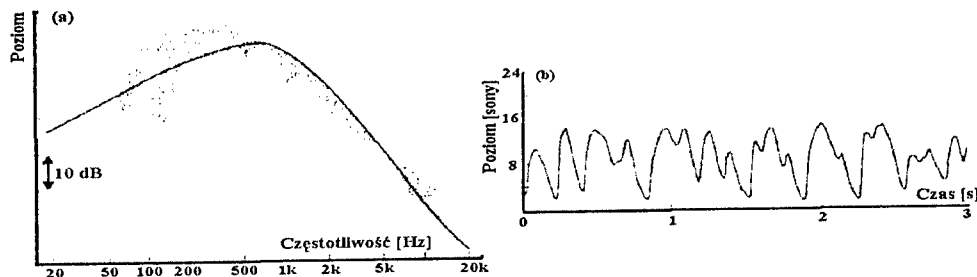
Celem tego testu jest ocena poprawności działania algorytmu dla sygnału mowy w rzeczywistych warunkach odsłuchowych. Podstawą oceny są subiektywne wrażenia ekspertów. Test realizowany jest w trzech krokach. W pierwszym etapie osoba badana zapoznaje się z instrukcją testową. Zawiera ona informację o problemie badawczym i ma na celu wskazanie kryteriów oceniania poszczególnych sygnałów w teście. W drugim etapie prezentowane są sygnały kontrolne umieszczone w pięciu kanałach systemu odsłuchowego [7]. Ta część testu ma na celu zapoznanie osoby badanej z poziomem

ciśnienia dźwięku generowanym przez głośniki. Dodatkowo spełnia ona funkcję prezentacji przykładowych sygnałów testowych. Jako sygnały kontrolne jak i późniejsze sygnały testowe zastosowano wypowiedzi lektorki. Trzeci etap testu polega na prezentacji piętnastu sygnałów poddanych normalizacji głośności. Sposób prezentacji poszczególnych przypadków wynika z wybranej metody porównywania. Metoda preferencji dwójkowych – 2WM (ang. *preference method*) polega na zestawieniu sygnałów w parach oraz oceny ich według kryterium „lepszy – gorszy”. Dzięki takiej metodzie można stwierdzić, który z prezentowanych przypadków spełnia w większym stopniu przyjęte założenia. Dotyczą one spójności panoramy dźwiękowej – czyli braku możliwości wskazania konkretnego położenia źródła dźwięku w takiej panoramie. Przyjęto, iż tak sformułowane założenia są równoznaczne z poprawnie wykonaną normalizacją głośności. Wybrana technika komparatystyczna to technika par niezależnych (RCSP ang. – *roving standard procedure*). Dany sygnał porównywany jest z sygnałem poprzednim i następnym (zasada A-B-A, gdzie A oznacza sytuację przed normalizacją głośności a B sytuację po normalizacji głośności). Zastosowany wzorzec porównań w dużym stopniu neutralizuje błąd czasu [10]. W teście wykorzystuje się trzystopniową skalę ocen o stopniach: „lepiej”, „bez zmian”, „gorzej”. Łętowski wskazuje, że zastosowanie trzystopniowej skali uwypukla przede wszystkim duże różnice pomiędzy prezentowanymi sygnałami. Skala dwustopniowa pozwala na stwierdzenie bardziej subtelnych różnic [10]. W świetle założeń dotyczących celu badań użycie trzystopniowej skali ocen jest bardziej zasadne. Dla ograniczenia czasu trwania testu zastosowano formę sekwencji wymuszonych. W konsekwencji czas trwania nie przekracza 5 minut, dzięki czemu w pełni wyeliminowano błąd czasu [10].

Prezentowane w dalszej części artykułu wyniki obejmują odpowiedzi ośmiu ekspertów. Niewielka liczba badanych osób nie pozwala na przeprowadzenie analizy statystycznej otrzymanych wyników.

W teście wykorzystywane są następujące sygnały: szum Fastla, ząb piły o częstotliwości 240Hz, szum pasmowy oraz nagrania lektora i lektorki. Drugi i trzeci z wymienionych sygnałów wybrano opierając się na analizie procesu artykulacji mowy. Mają one za zadanie symulować (zgrubnie) parametry tonu krtaniowego (ząb piły) i głosek szumowych (szum pasmowy). Opis wspomnianego procesu można znaleźć w książce Czesława Basztury [11].

Szum Fastla został wybrany ponieważ jego widmo i przebieg czasowy odzwierciedlają odpowiednie parametry płynnej mowy polskiej [12]. Sygnał ten uzyskuje się poprzez modulację amplitudową znanego z audiologii szumu CCITT [13]. Widma obu sygnałów pozostają identyczne. Rysunek 7 przedstawia widmo (a) i przykładowy przebieg czasowy szumu Fastla (b) [14].



Rys.7. (a) – widmo szumów CCITT i Fastla, przykładowy przebieg czasowy szumu Fastla – (b) [14].

Wybór nagrań lektorskich podyktowany jest głównym celem testu (zbadanie poprawności działania algorytmu dla ludzkiej mowy)

W dalszej części artykułu prezentowane są otrzymane wyniki. Tabela 3.1 prezentuje liczby odpowiedzi poszczególnych ekspertów.

Tablica 3.1

Sumaryczna liczba odpowiedzi poszczególnych ekspertów

Ekspert	Liczba odpowiedzi		
	„lepiej”	„bez zmian”	„gorzej”
1	15	0	0
2	9	4	2
3	7	8	0
4	8	5	2
5	12	3	0
6	15	0	0
7	8	7	0
8	12	3	0
Średnia	10,75	3,75	0,5

W tabeli 3.2 prezentowane są wyskalowane w procentach sumy oceny dla poszczególnych sygnałów testowych uzyskane od wszystkich badanych ekspertów

Tablica 3.2

Liczba ocen (procentowo) ekspertów dla poszczególnych badanych sygnałów

Sygnał	„lepiej”	„bez zmian”	„gorzej”
Szum Fastla	62,5	37,5	0
Ząb piły	62,5	29,2	8,3
Szum pasmowy	79,2	16,7	4,1
Głos męski	87,5	12,5	0
Głos żeński	75	20,8	4,2

Wyniki pokazują poprawność pracy algorytmu szczególnie dla głosu lektora. Również głos lektorki i szum pasmowy symulujący głoski szumowe, są przetwarzane poprawnie. Nieco gorsze wyniki uzyskano dla dwóch pozostałych sygnałów syntetycznych (szum Fastla i ząb piły). Sumaryczna liczba odpowiedzi prezentowana w tabeli 3.1 świadczy o dobrej skuteczności algorytmu.

6. WNIOSKI

W artykule przedstawiono opis algorytmu normalizacji głośności bazujący na wartości skutecznej sygnałów. Algorytm wykorzystuje funkcje *HRIR*. Praktyczna realizacja została wykonana w środowisku *Matlab*. Na jej podstawie przeprowadzono szereg prostych testów mających na celu potwierdzenie poprawności działania algorytmu. Pilotażowy charakter

eksperymentów (w przypadku pierwszego z opisywanych testów – podpunkt 3.2) i niewielka liczba badanych osób (w przypadku drugiego z opisywanych testów – podpunkt 3.3) nie pozwalają na wyciągnięcie daleko idących wniosków. Nie mniej jednak otrzymane wyniki wskazują na możliwość zastosowania opisanego rozwiązania w prostych systemach przetwarzania dźwięku dedykowanym sygnałom mowy.

PODZIĘKOWANIE

Praca została dofinansowana ze środków projektu celowego nr 113/BO/B.

BIBLIOGRAFIA

- [1] Glasberg R. B., Moore C. J.: *A model of loudness applicable to time-varying sounds*, J. Audio Eng. Soc., Vol. 50, No. 5, s. 331-342, May 2001.
- [2] Glasberg R. B., Moore C. J., Baer T.: *A model for the prediction of threshold, loudness, and partial loudness*, J. Audio Eng. Soc., Vol. 45, No. 4, s. 224-240, April 1997.
- [3] Maziewski, P.: *System przetwarzania sygnałów do celu odbioru dźwięku wielokanałowego za pomocą słuchawek stereofonicznych*, X Sympozjum Inżynierii i Reżyserii Dźwięku, Wrocław 2003.
- [4] Maziewski P.: *Technika wirtualizacji wykorzystująca odpowiedzi impulsowe zarejestrowane za pomocą sztucznej głowy w komorze bezechowej*, PTETiS, Gdańsk 2003.
- [5] Zwickler E.: *Procedure for calculating loudness of temporally variable sounds*, J. Acoust. Soc. Am., Vol. 62, s. 675-682, 1997.
- [6] Keith M., Gardner B.: *HRTF measurements of a kemar dummy-head microphone*, Report 280, MIT Media Lab, May 1994.
- [7] ITU-R: Recommendation BS. 775-1: *Multichannel stereophonic sound system with and without accompanying picture*, Geneva 1997
- [8] Zwickler E., Fastl H.: *Psychoacoustics Facts & Models*, Springer, Berlin 1990.
- [9] Dziubiński M., *Skrypty służące go generacji krzywych głośności*, KSM PG, Gdańsk 2003.
- [10] Łętowski, T.: *Słuchowa ocena sygnałów i urządzeń*, Akademia Muzyczna w Warszawie, Warszawa 1984.
- [11] Basztura Cz.: *Rozmawiać z komputerem*, Wydawnictwo Prac Naukowych "FORMAT", Wrocław 1992
- [12] Tarnoczy T.: *Das durchschnittliche Energie-Spektrum der Sprache für sechs Sprachen*, Acustica, Vol. 24, s.56-74, 1971.
- [13] ITU-T: Recommendation G.227: *Conventional Telephone Signal*
- [14] Hojan E., Fastl H.: *Intelligibility of Polish and German speech for the Polish audience in the presence of noise*, Archives of Acoustica, Vol. 21, No 2, s. 123-130, 1996.

LOUDNESS NORMALIZATION ALGORITHM FOR SOUNDFILES

Summary

An algorithm for sound files loudness normalization is shown. It is dedicated for use with a surround loudspeaker system. The algorithm generates values of gains necessary to normalize loudness of chosen sound files. This is achieved by *rms* values normalization preceded by the HRTF (Head Related Transfer Function) base filtering. The last step of procedure is a maximum value normalization, necessary to eliminate data clipping. Results of two tests are shown. Additionally a short overview of loudness perception models, which use objective values to represent subjective loudness, is also included.

Renata Kalicka

Katedra Inżynierii Biomedycznej, Politechnika Gdańska

MODELOWANIE PERFUZJI MÓZGU W BADANIACH MRI

Streszczenie

Praca dotyczy dynamicznego badania mózgu z wykorzystaniem techniki MRI (badanie perfuzji). W rezultacie pomiarów, tj. czasowej sekwencji skanów MRI w odpowiedzi na podanie znacznika zmieniającego właściwości magnetyczne tkanek, wyznaczony zostaje model procesu perfuzji. Dla badanego procesu perfuzji opracowano model nieparametryczny (odpowiedź aproksymowana niezupełną funkcją gamma o 3 parametrach) oraz dwukompartментowy model parametryczny. W przypadku modelu parametrycznego, zaproponowano analityczny opis funkcjonowania systemu w postaci prawdopodobnej jego struktury i zbioru jej parametrów (mikroparametrów modelu). Oba modele zostały przebadane (z wykorzystaniem programu Matematica) i porównane.

1. WSTĘP

Obserwuje się znaczący wzrost ilości chorób mózgu, szczególnie w grupie chorób ościennych (Alzheimer), demielinizacyjnych i nowotworowych. Wskazuje to na potrzebę doskonalenia metod wczesnej diagnostyki i terapii. Stosowane dotychczas metody, łączące techniki badań funkcjonalnych i strukturalnych, okazują się niewystarczające. Podkreśla się potrzebę wdrożenia metod analizy dynamicznej oraz integracji badań metodami PET i MRI. Celem jest pozyskanie, w rezultacie badań dynamicznych i modelowania, tzw. obrazów parametrycznych. Niniejsza praca dotyczy badania dynamicznego mózgu z wykorzystaniem techniki MRI. Na podstawie zgromadzonej czasowej sekwencji skanów MRI w odpowiedzi na podanie znacznika zmieniającego właściwości magnetyczne tkanek, wyznacza się model procesu perfuzji. Parametry tego modelu, np. stałe czasowe dystrybucji i eliminacji znacznika, zostaną wykorzystane do przestrzennego zobrazowania zmian parametru w obranym obszarze mózgu. Tego typu zobrazowania parametryczne mają większą wartość diagnostyczną niż przedstawia dowolny ze skanów MRI.

Dla procesu perfuzji opracowany został model nieparametryczny (odpowiedź aproksymowana niezupełną funkcją gamma o 3 parametrach) oraz dwukompartментowy model parametryczny. W przypadku wykorzystania do modelowania funkcji gamma [1,2], celem jest dobranie parametrów tej funkcji (makroparametrów modelu), tak aby wiernie naśladowała mierzone zmiany perfuzji. Utworzono także model parametryczny; zaproponowano opis funkcjonowania systemu w postaci prawdopodobnej jego struktury i zbioru para-

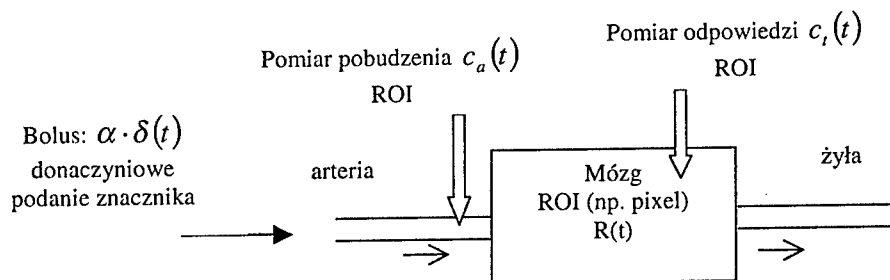
metrów (mikroparametrów modelu). Oba modele zostały przebadane (z wykorzystaniem programu Matematica) i porównane.

2. MIARY WŁAŚCIWOŚCI CHEMODYNAMICZNYCH BADANYCH OBSZARÓW MÓZGU

Ocena właściwości hemodynamicznych dotyczy stanu naczyń i związanej z tym jakości krążenia krwi. Perfuzja, czyli stopień ukrwienia tkanek, jest jednym z najważniejszych wskaźników żywotności i funkcjonalności tkanek [3-5]. Do konwencjonalnych pomiarów perfuzji wykorzystuje się angiografię rentgenowską, tomografię komputerową poprzedzoną inhalacją stabilnego ksenonu oraz tomografię PET. Gdy dostępne stały się techniki szybkiego zobrazowania MRI, stało się możliwe zastosowanie ich do nieinwazyjnych pomiarów przepływu krwi mózgowej. W tym celu stosuje się znacznik (paramagnetyk, najczęściej pochodną gadoliny, rzadziej manganu) umożliwiający pomiar perfuzji z wykorzystaniem dynamicznych badań MRI. Przykładem jest podanie bolusa znacznika Gd-DTPA (gadolina) i następnie śledzenie T2 zależnych zobrazowań MR. Gd-DTPA jest kontrastem hydrofilnym (wodochłonnym). Obszarem jego dystrybucji jest przestrzeń wewnątrznacyniowa i zewnątrzkomórkowa. Podany dożylnie jest stosunkowo szybko wydalany w postaci niezmienionej przez filtrację nerkową. Nie przechodzi przez prawidłową barierę krew-mózg, pozostaje w naczyniach. Jeśli po podaniu naczyniowym obserwuje się jego obecność w tkance mózgowej, oznacza to zawsze patologię, uszkodzenie bariery krew-mózg. Wewnątrz mózgowo może pojawić się tylko w strukturach, które nie mają bariery krew-mózg, tj. w oponach, w przysadce i w splotach naczyniówkowych.

Znacznik Gd-DTPA, zmieniający podatność magnetyczną tkanek, podany w postaci bolusa do krwiobiegu, umożliwia śledzenie i zobrazowanie dynamiki jego zmian w ROI (*Region Of Interest*). Do modelowania i do obliczeń wykorzystywane są dane pomiarowe z pierwszego przejścia znacznika przez ROI; jest to tzw. *first pass fit*. Pomiary i obliczenia służą do wyznaczenia map przepływu krwi mózgowej (CBF – *Cerebral Blood Flow*), map objętości krwi mózgowej (CBV – *Cerebral Blood Volume*) oraz map średniego czasu przejścia (MTT – *Mean Transit Time*).

Przedstawmy ROI w mózgu wraz z naczyniem zasilającym ten obszar w krew oraz naczyniem, którym krew opuszcza ten obszar:



Rys.1. Pobudzenie i odpowiedź ROI w badaniach perfuzji metodą MRI z wykorzystaniem znacznika zmieniającego podatność magnetyczną tkanek.

Wielkością wejściową dla ROI jest $c_a(t)$ (*arterial input function*), mierzone możliwie blisko pixela (w dużym naczyniu, czyli we krwi). Następnie obserwujemy pixel. Odpowiedzią jest to, co pozostaje w pixelu, to co w nim rezyduje (*residue function*) $R(t)$. Zatem $R(t)$ jest odpowiedzią impulsową mierzoną we wnętrzu pixela, a nie na jego wyjściu. Dla odpowiedzi impulsowej $R(t)$ oraz dla pobudzenia $c_a(t)$, aktualna odpowiedź $c_t(t)$ (*flow arriving pixel*) wewnątrz pixela (w małych naczyniach, w tkankach) wyrażona jest przez spłot:

$$c_t(t) = \int_0^t c_a(\tau) \cdot FR(t - \tau) d\tau = c_a(t) \otimes FR(t) \quad (2.1)$$

F to względny strumień krwi wpływający do pixela. Ilość znacznika jaka trafi do badanego pixela $r=1,2,\dots,n$ jest proporcjonalna do części strumienia wejściowego F_{in}

niosącego masę m_{in} : $F_{in} m_{in} = \sum_{r=1}^n F_r m_r$.

Sekwencja czasowa skanów MR pixela daje pomiar $c_t(t)$ dla ROI. Skanujemy także $c_a(t)$, więc stosując rozplot, wyznaczymy $FR(t)$. To z kolei pozwala wyznaczyć MTT , CBV i CBF jako względny strumień F dopływający do pixela.

Wielkość MTT wyznacza się jako wartość oczekiwaną \bar{t} zmiennej losowej t , traktując $h(t)$ jako gęstość prawdopodobieństwa tej zmiennej losowej:

$$MTT = \bar{t} \stackrel{def}{=} 1^{st} \text{ moment of } h(t) = \int_0^{\infty} t \cdot h(t) dt \quad (2.2)$$

Wykorzystując związek odpowiedzi impulsowej $h(t)$ na wyjściu pixela, z odpowiedzią

w jego wnętrzu $R(t)$: $R(t) = 1 - \int_0^t h(\tau) d\tau$ otrzymujemy: $MTT = \int_0^{\infty} R(t) dt$. Po rozplocie

mamy wielkość $FR(t)$, tak więc praktycznie możemy wyznaczyć:

$$MTT = \bar{t} = \int_0^{\infty} FR(t) dt \quad (2.3)$$

Następny z poszukiwanych parametrów to CBV . Wyznacza się go z definicji:

$$CBV = \frac{\int_0^{\infty} c_t(t) dt}{k \cdot \int_0^{\infty} c_a(t) dt} = \frac{\text{liczba proporcjonalna do } V \text{ pixela}}{\text{liczba proporcjonalna do } V \text{ systemu}} \quad (2.4)$$

Licznik powyższego wyrażenia jest interpretowany jako miara liczby cząstek znacznika, które osiągnęły określony ROI. Ta liczba jest proporcjonalna do względnej objętości badanego ROI. W mianowniku znajduje się wielkość proporcjonalna do całkowitej ilości znacznika wprowadzonego do systemu, więc proporcjonalna do całkowitej jego objętości. Tak więc, CBV dane wzorem (2.4) jest miarą względnej objętości ROI. Powyższe rozumowanie jest słuszne tylko wtedy, gdy nie jest uszkodzona bariera krew-mózg.

3. MODELOWANIE ORAZ IDENTYFIKACJA MODELI SYSTEMÓW FIZJOLOGICZNYCH

Model badanego systemu fizjologicznego przedstawia się w postaci połączonych wzajemnie podsystemów, z których każdy jest opisany równaniem różniczkowym, lub jego odpowiednikiem. Najczęściej stawianym problemem praktycznym jest zagadnienie identyfikacji, tzn. zagadnienie poszukiwanie estymat parametrów modelu.

W ogólności możliwe są dwa podejścia do identyfikacji:

1. Gdy system jest słabo poznany i brak podstaw do sformułowania r.r., które przybliżają działanie systemu, stosujemy **podejście nieparametryczne**. Poszukujemy odpowiedzi impulsowej $h(t)$ systemu: $y(t) = h(t) \Big|_{u(t)=\delta(t)}$.

Na tej podstawie ocenia się właściwości systemu. Rezultat jest czysto empiryczny.

2. Gdy dysponujemy pewną wiedzą o systemie, stosujemy **podejście parametryczne**. Postulujemy pewną strukturę modelu, tj. układ równań wraz ze zbiorem parametrów, wciąż jednak wartości liczbowe parametrów nie są znane. Struktura wymaga identyfikacji parametrów.

Wśród nieparametrycznych metod identyfikacji stosowane są: numeryczny rozplot (numeryczna dekonwolucja), estymacja odpowiedzi impulsowej z wykorzystaniem funkcji kowariacyjnych oraz estymacja odpowiedzi impulsowej w dziedzinie częstotliwości.

Identyfikacja modelu parametrycznego przebiega w dwu etapach:

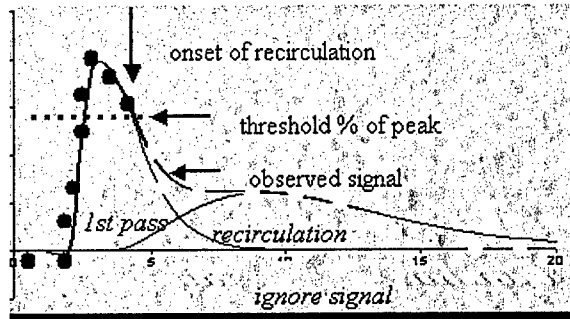
1. Sformułowanie hipotezy o prawdopodobnej zależności wejście-wyjście, na ogół w postaci równań różniczkowych, całkowych, algebraicznych. Szczególnie przydatny jest tu opis systemu w kategoriach zmiennych stanu.
2. Estymacja nieznanych parametrów przez minimalizację obranego kryterium J obrazującego jakość dopasowania odpowiedzi modelu $y_{mod}(\mathbf{p}, nT)$ i zmierzonej odpowiedzi systemu $y(nT)$, np. w postaci:

$$J = \sum_{n=0}^{N-1} [y(nT) - y_{mod}(\mathbf{p}, nT)]^2 = \sum_{n=0}^{N-1} e^2(nT) = \min.$$

Jeśli istnieją podstawy do sformułowania hipotezy o mechanizmie działania systemu, należy wybrać modelowanie parametryczne, gdyż przybliża nas ono do poznania mechanizmów funkcjonowania sytemu, w odróżnieniu od modelowania nieparametrycznego, którego celem jest jedynie znalezienie wiernego opisu odpowiedzi systemu.

4. MODEL PARAMETRYCZNY I MODEL NIEPARAMETRYCZNY PERFUZJI MÓZGU W BADANIACH MRI

Do modelowania i obliczeń wykorzystywane są tzw. dane pomiarowe z pierwszego przejścia znacznika przez ROI, jest to tzw. *first pass fit*. Dla przykładowego zbioru danych pomiarowych [7] opracowano model nieparametryczny a następnie model parametryczny.



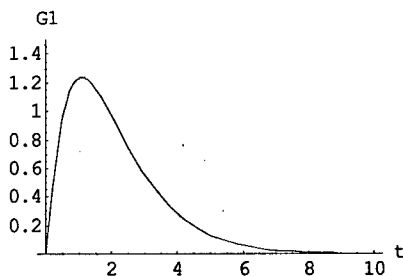
Na powyższym wykresie przedstawione są przykładowe dane pomiarowe w badaniach perfuzji mózgu. W sygnale mierzonym (*observed signal*) przedmiotem zainteresowania jest część pomiarów odnosząca się do *first pass*. Część odpowiedzi na którą wpływa recyrkulacja jest ignorowana w badaniach perfuzji mózgu.

Modelowanie nieparametryczne black-box

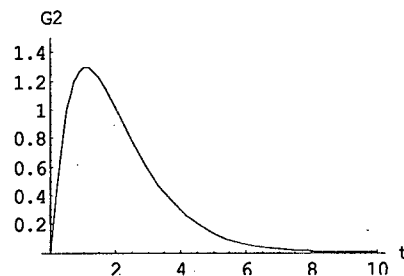
Dla rozważanego przypadku modelowania perfuzji, odpowiedź impulsową aproksymuje się [1-3] za pomocą uogólnionej funkcji gamma $Gama[a, z_0, z_1] = \int_{z_0}^{z_1} e^{-t} t^{v-1} dt$, którą

można wyrazić [6] w postaci: $\int e^{gt} t^{-n} dt = \frac{1}{n-1} \left[-e^{gt} t^{-(n-1)} + g \int e^{gt} t^{-(n-1)} dt \right]$.

Wyrażenie funkcji *Gama* za pomocą jednego składnika sumy daje funkcje regresji $G1[a, b] = ae^{-t} t^b$.



Rys.2. Wykres funkcji G1
dla $0 \leq t \leq 10$ oraz dla
 $a = 3.35$, $b = 1.10$.

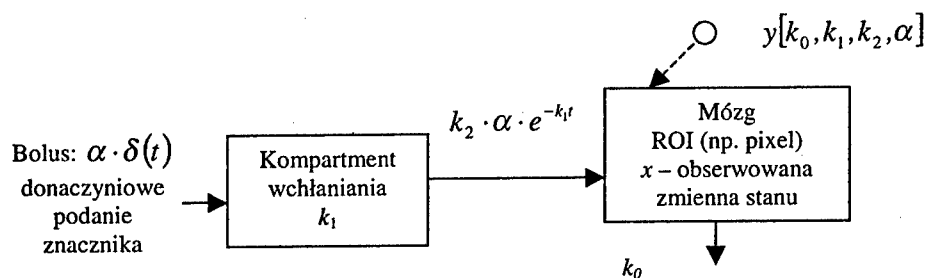


Rys.3. Wykres funkcji G2
dla $0 \leq t \leq 10$ oraz dla
 $c = 0.025$, $b = 1.100$, $d = 150$.

Funkcja regresji z rys. 2 ma oczekiwany kształt, lecz wykazuje zbyt małą elastyczność postaci przy zmianie wartości parametrów. Oznacza to w praktyce trudności w naśladowaniu zmian wielkości mierzonej. Dla dwóch składników sumy funkcja regresji przyjmuje postać $G2[b, c, d] = ce^{-t}t^b(td - 1)$. Funkcja z rys. 3 prezentuje typ zmienności podobny do G1, a wprowadzenie trzeciego parametru bardzo poprawia elastyczność opisu. Funkcja regresji G2 zostaje przyjęta jako odpowiedź modelu *black-box*.

Modelowanie parametryczne – model kompartmentowy

Kompartментowy model dystrybucji znacznika przedstawiony jest na rys. 4. Wyróżniono kompartment wchłaniania i kompartment mózgowy.



Rys.4. Kompartментowy model dystrybucji znacznika.

Zmienna stanu x jest obserwowana jako odpowiedź modelu $y[k_0, k_1, k_2, \alpha]$, gdzie k_0, k_1, k_2 to mikroparametry modelu, α oznacza ilość wprowadzonego do krwiobiegu znacznika.

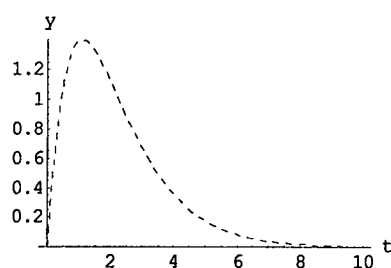
Model opisany jest w następujący sposób:

$$\begin{aligned} \dot{x} &= k_2 \cdot \alpha \cdot e^{-k_1 t} - x \cdot k_0, \quad x(0) = 0, \\ y &= x \end{aligned} \quad (4.1)$$

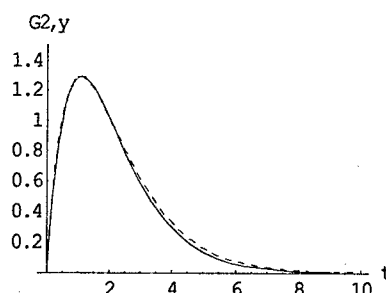
Rozwiązanie powyższego równania ma postać:

$$y = \frac{k_2 \cdot \alpha}{k_0 - k_1} (e^{-k_1 t} - e^{-k_0 t}) \quad (4.2)$$

Wykres odpowiedzi modelu kompartmentowego (4.2) dla zbioru mikroparametrów oszacowanych na podstawie doniesień literaturowych dotyczących perfuzji mózgu [2,4] przedstawiono na rys. 5.



Rys.5. Wykres odpowiedzi modelu y dla $0 \leq t \leq 10$ oraz dla $k_0 = 0.85 [1/h]$, $k_1 = 1.00 [1/h]$, $k_2 = 1.85 [1/h]$ oraz $\alpha = 1.90$.



Rys.6. Porównanie funkcji regresji $G2$ oraz y

Wartości parametrów k_0, k_1, k_2 obrano, jako prawdopodobne, na podstawie literatury. Należy dokonać wyboru pomiędzy funkcją $G2$ dla modelu *black-box* oraz odpowiedzią y modelu parametrycznego. Jak wynika z rys. 6 są one niemal nierozróżnialne graficznie.

5. ZAKOŃCZENIE

Uzyskane rezultaty pokazują, że wierność odwzorowania wyników pomiarów przez funkcje regresji obu modeli – $G2$ dla nieparametrycznego *black-box* oraz y dla parametrycznego modelu kompartmentowego – są takie same. Badany model parametryczny i nieparametryczny wykazują taką samą elastyczność w zdolności naśladowania danych pomiarowych. Model parametryczny, dwukompartmentowy, daje dodatkowo wgląd w sposób funkcjonowania systemu, daje możliwość oceny np. szybkości procesów dystrybucji i eliminacji, a więc możliwość utworzenia zobrazowania tych stałych czasowych w badanym obszarze mózgu. Ma to bardzo dużą wartość diagnostyczną, gdyż może wskazać na niejednorodności w tkance mózgowej, które mogą wystąpić znacznie wcześniej niż pojawiają się jakiegokolwiek symptomy chorobowe.

BIBLIOGRAFIA

- [1] T.-Q. Li, Z. G. Chen, L. Ostergard, T. Hindmarsh, M. E. Mosley: *Quantification of cerebral blood flow by bolus tracking and artery spin tagging methods*, Magnetic Resonance Imaging 18, pp. 503-512, Elsevier Science 2000.
- [2] F. Calamante, D.G. Gadian, A. Connelly: *Quantification of Perfusion Using Bolus Tracking Magnetic Resonance Imaging in Stroke*. Comments, Opinions and Reviews, January 2003, pp.1146-1151, American Heart Association, Inc.
- [3] A. Klienschmidt, H. Steinmetz, M. Sitzler, et al.: *Magnetic Resonance Imaging of Regional Cerebral Blood Oxygenation Changes Under Acetazolamine in Carotid Occlusive Disease*. Stroke 1995;26:106-10.
- [4] A.G. Sorensen, W.A. Copen, L. Ostergaard: *Hyperacute Stroke: Simultaneous Measurement Of Relative Cerebral Blood Volume, cerebral Blood Flow and Mean Tissue Transit Time*. Radiology 1999;210:512-27
- [5] R.R. Edelman, H.P. Mattle, D.J. Atkinson et al.: *Cerebral Blood Flow: Assessment with Dynamic Contrast Enhanced T*2-weighted MR Imaging at 1.5T*. Radiology 1990;176:211-20.
- [6] G.A. Korn, T.M. Korn: *Matematyka dla pracowników naukowych*, cz.2, PWN, 1983.
- [7] Perfusion Analysis in MEDx3.3 Sensor systems, INC.

MODELING OF PERFUSION IN DYNAMIC MRI INVESTIGATION OF A BRAIN

Summary

This study deals with modeling of perfusion. The term perfusion refers to the delivery of blood at the level of the capillaries, where exchange of oxygen and nutrients between blood and tissue takes place. MR techniques have been very powerful in providing indicators of tissue perfusion in the brain. Bolus injection of exogenous tracers such as Gd-DTPA has been used for first-pass bolus tracking imaging. Modeling of the process and tracer kinetic analysis allow to create maps of CBV, MTT, CBF as well as, in a case of parametric modeling, maps of microparameters of the system under investigation.

Wiesław Paja

Katedra Systemów Ekspertowych i Sztucznej Inteligencji, Wyższa Szkoła Informatyki
i Zarządzania z siedzibą w Rzeszowie

INTERNETOWY SYSTEM DIAGNOZOWANIA ZMIAN MELANOCYTOWYCH SKÓRY

Streszczenie

W pracy przedstawiono praktyczną realizację internetowego systemu diagnozowania zmian melanocytowych skóry (czerniaka). System ten posiadający również wyraźne funkcje dydaktyczne, oblicza wartości parametru **TDS** (Total Dermatoscopy Score) posługując się trzema odmiennymi algorytmami. Pierwszy z nich realizuje obliczenia wg klasycznej formuły Braun-Falco (tradycyjny **TDS**), drugi sposób wykorzystuje zoptymalizowaną formułę (obliczany jest **nowy TDS**), zaś trzeci algorytm wykorzystuje opracowane drzewo decyzji. Opracowany system pozwala na wczesną, nieinwazyjną diagnozę zmiany skórnej. Formą pomocy jest zestaw instrukcji pozwalających na zrozumienie zasad określania i/lub obliczania parametrów wskaźnika **TDS**. Zawiera on omówienie diagnozy autentycznych przypadków zmian skóry. Przedstawiony system jest dostępny w sieci Internet pod adresem <http://www.wsiz.rzeszow.pl/ksesi>.

1. WSTĘP

Głównym celem projektu był zamysł wykorzystania własnej bazy informacyjnej do badań dotyczących zastosowania metod uczenia maszynowego w ocenie hierarchii ważności symptomów zmian melanocytowych i czerniaka skóry, wchodzących w definicję powszechnie stosowanego wskaźnika zagrożenia rozpatrywanym nowotworem. Wskaźnik ten, znany pod skrótem **TDS** (Total Dermatoscopy Score, wg Braun-Falco [1]), jest obliczany z następującej zależności

$$\text{TDS} = 1.3 * \text{Asymetria} + 0.1 * \text{Brzeg} + 0.5 * \Sigma \text{Kolory} + 0.5 * \Sigma \text{Struktura} \quad (1.1)$$

w której brane są pod uwagę następujące rodzaje symptomów (zwanymi dalej *atrybutami opisującymi*):

<Asymetria> przyjmuje trzy wartości logiczne: *asymetria 1-osiowa*, *asymetria 2-osiowa*, *zmiana symetryczna*

<Brzeg> atrybut numeryczny, przyjmuje wartości 0 – 8

<**Kolor**> przyjmuje sześć wartości: *biały, błękitny, ciemny_brąz, jasny_brąz, czarny, czerwony*

<**Struktura**> przyjmuje pięć wartości: *ciałka barwnikowe, kropki barwnikowe, pole bezstrukturalne, rozgałęzienia pasmowe, sieć barwnikowa*

Na podstawie wartości wymienionych cech (atrybutów) rozpoznaje się cztery rodzaje znamion melanocytowych, a mianowicie: *znamię łagodne, znamię błękitne, znamię podejrzanе oraz znamię złośliwe*.

2. BAZA PRZYPADKÓW

W badaniach wykorzystano bazę danych przypadków medycznych zgromadzonych w Wojewódzkiej Przychodni Dermatologicznej w Rzeszowie. Powstała ona w wyniku badań rejestrowanych w postaci odręcznych opisów (historia choroby) wykonywanych przez lekarzy. Realizowano je w formie ankiety, w której stan pacjenta określono w postaci cech niezbędnych do obliczenia wskaźnika TDS. Na podstawie anonimowych kopii wspomnianych ankiet, w Katedrze Systemów Ekspertowych i Sztucznej Inteligencji Wyższej Szkoły Informatyki i Zarządzania w Rzeszowie, została utworzona baza danych wykorzystywana w badaniach. Baza zawiera 548 przypadków pacjentów posiadających znamiona melanocytowe widoczne na skórze [2]. Każdy z nich opisany jest przy pomocy 13 atrybutów opisujących przedstawionych w rozdziale poprzednim, oraz atrybutem czternastym, jakim jest wskaźnik TDS, wyliczanym na podstawie wspomnianych 13 atrybutów. Do obliczenia TDS stosujemy zależność (1.1). W ten sposób została utworzona możliwość weryfikacji badań przy pomocy koncepcji *konstruktywnej indukcji* [3] w uczeniu maszynowym. Każdy z przypadków jest przypisany do jednej z czterech klas decyzyjnych (jest zdiagnozowany jako jeden z czterech przypadków zmian melanocytowych skóry): *znamię łagodne, znamię błękitne, znamię podejrzanе lub znamię złośliwe*.

3. STRUKTURA I ZASADA DZIAŁANIA SYSTEMU

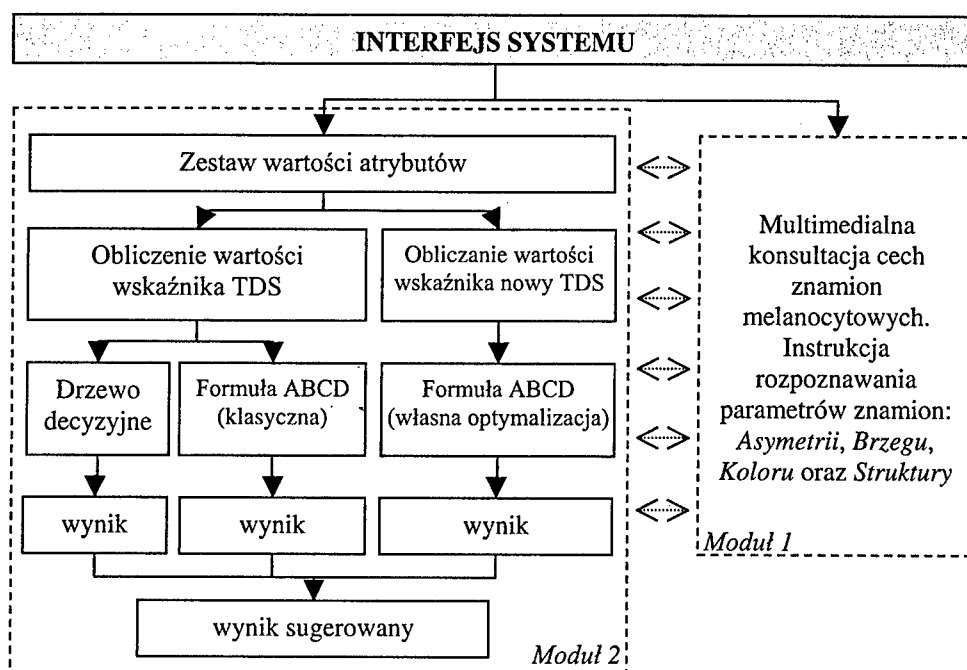
System posiada interfejs w postaci serwisu internetowego umożliwiającego dostęp do kolejnych modułów systemu. W strukturze systemu można wyróżnić dwa główne moduły. Pierwszy z nich to moduł multimedialny służący do wirtualnej konsultacji cech znamion melanocytowych na skórze. Pełni on rolę swoistego rodzaju instrukcji wczesnego rozpoznawania i wyznaczania parametrów charakteryzujących schorzenie. Parametry te to wspomniane wyżej *Asymetria* znamienia, *Brzeg* znamienia, *Kolor* znamienia oraz *Struktura* widoczna na znamieniu.

Asymetria – jest miarą asymetrii znamienia, z wartościami 0, 1, 2, odpowiednio dla zmiany symetrycznej (posiadającej poziomą i pionową oś symetrii), jednej osi symetrii (asymetria 1-osiowa) lub braku jakiejkolwiek osi symetrii (asymetria 2-osiowa),

Brzeg – jest charakterystyką brzegu znamienia, z wartościami od 0 do 8, wskazującymi liczbę ósmych części obwiedni znamienia, w których występuje ostre odcięcie pigmentowej frakcji znamienia skóry,

Kolor – jest liczbą (od 1 do 6) zdefiniowanych kolorów (*biały, błękitnoszary, ciemnobrązowy, jasnobrązowy, czarny, czerwony*), występujących jednocześnie w rozpatrywanym znamieniu,

Struktura – jest liczbą (od 1 do 5) zdefiniowanych rodzajów struktury znamienia, przypisując po 1 za obecność każdej ze zdefiniowanych struktur (*sieć pigmentowa, kropki barwnikowe, globulki barwnikowe, rozgałęzienia pasmowe, obszary bezstrukturalne*), występujących jednocześnie w analizowanym przypadku schorzenia.



Rys. 1. Struktura systemu

Korzystając z pierwszego modułu użytkownik może poznać prawidłowe metody diagnozowania poszczególnych parametrów znamienia na przykładzie prawdziwych przypadków schorzeń.

Moduł drugi jest rodzajem kalkulatora melanocytowego umożliwiającego nieinwazyjną diagnozę zmian melanocytowych. Wiadomo, że poprawna klasyfikacja znamion barwnikowych skóry jest możliwa jedynie na podstawie badania histopatologicznego. Niezmienny trend nowoczesnej diagnostyki, a mianowicie dążenie do używania metod nieinwazyjnych, stał się przyczyną upowszechnienia, jako kryterium oceny stopnia zagrożenia czerniakiem, wzmiankowanego wskaźnika TDS, mimo że w niektórych kręgach lekarskich jest on poddawany krytyce.

Wartościami wejściowymi do modułu drugiego jest zestaw wartości 13 atrybutów opisujących dany przypadek schorzenia. Wartości te wprowadzone przez użytkownika systemu poprzez odpowiedni formularz są przetwarzane przy użyciu dwóch zależności obliczających 14 atrybut, jakim jest wskaźnik TDS. Pierwsza z nich to standardowa reguła ABCD przedstawiona wcześniej w równaniu 1.1, natomiast druga jest to zależność ze współczynnikami zoptymalizowanymi [4,5], pozwalająca wyznaczyć **nowy_TDS** szczegółowo przedstawiony w [4], mający postać:

$$\begin{aligned} \text{nowy_TDS} = & (0.8 * \text{Asymetria}) + (0.11 * \text{Brzeg}) + (0.5 * \text{kolor_biały}) + \\ & (0.8 * \text{kolor_błękitnoszary}) + (0.5 * \text{kolor_ciemnobrązowy}) + \\ & (0.6 * \text{kolor_jasnobrązowy}) + (0.5 * \text{kolor_czarny}) + (0.5 * \text{kolor_czerwony}) + \\ & (0.5 * \text{sieć_pigmentowa}) + (0.5 * \text{kropki_barwnikowe}) + (0.6 * \text{globulki_barwnikowe}) + \\ & (0.6 * \text{rozgałęzienia_pasmowe}) + (0.6 * \text{obszary_bezstrukturalne}) \end{aligned} \quad (1.2)$$

Wspomaganie diagnostyki zmian melanocytowych oraz czerniaka skóry

Asymetria:

Brzeg:

Barwa znamienia:

☐ jasny brąz ☒ ciemny brąz

☒ biały ☐ czerwony

☐ czarny ☒ szarymnięski

Struktura znamienia:

☒ ciążka ☐ rozgałęzienia

☐ barwnikowe ☐ pasmowate

☒ kropki ☐ sieć barwnikowa

☒ pola bezstrukturalne

DIAGNOZUJ **WYCZYŚĆ**

Współczynnik TDS

Nowy TDS

Przykłady diagnozowania asymetrii znamienia

Cechę Asymetria zmian melanocytowych definiujemy poprzez poszukiwanie jej symetrii. Po najechaniu kursorem na obraz zmiany i kliknięciu lewym klawiszem myszki, program uruchamia poszukiwanie pierwszej (głównej) osi symetrii, dzielącej zmianę na dwie części, symetryczne w kontekście brzegu, koloru i struktury. Użytkownik jednak powinien sam ocenić, czy utworzona oś faktycznie jest osią symetrii. Następnie program sprawdza, czy jest możliwe utworzenie drugiej osi symetrii, prostopadłej do osi pierwszej.

Asymetria: zmianę symetryczną

Asymetria: asymetria 1-osiowa

Asymetria: asymetria 2-osiowa

Przykłady diagnozowania brzegu znamienia

Przykłady diagnozowania barwy znamienia

Przykłady diagnozowania struktury znamienia

Formuła ABCD (klasyczna): **Formuła ABCD (własna optymalizacja):** **Drzewo decyzji:** **System sugeruje:**

Rys. 2. Interfejs systemu

Uzyskane wartości atrybutu TDS są podstawą do wystawienia diagnozy przy użyciu klasycznej Formuły ABCD gdzie zgodnie, z tzw. drugim algorytmem Stolca [6] (o złączonych wartościach progów klasyfikacji), a także wg innych źródeł [7], przyjmuje się, że:

TDS < 4.76 wskazuje **łagodną** zmianę melanocytową,
 4.76 ≤ TDS ≤ 5.45 wskazuje **podejrzaną** zmianę melanocytową,
 TDS > 5.45 wskazuje zmianę **złośliwą** (czerniaka skóry).

Analogiczne progi klasyfikacyjne obowiązują w przypadku Formuły ABCD z własną optymalizacją wskaźnika TDS (**nowy_TDS**).

Trzecią metodą diagnozowania jest wykorzystanie drzewa decyzyjnego. Drzewo decyzyjne uzyskano na podstawie bazy danych przedstawionej w paragrafie 2. Algorytm wykorzystany do utworzenia drzewa to algorytm R. Quinlana ID3/C4.5 [8]. W wyniku zastosowania takiego algorytmu otrzymano drzewo postaci:

```
TDS ??
<= 4.8 : C_BLEKITNY ??
      jest : ---> Mel_znmblek
      nie_ma : ---> Mel_lagodna
> 4.8 : TDS ??
      <= 5.3 : ---> Mel_podejrz
      > 5.3 : ---> Mel_zlosliw
```

Ostatecznie system podejmuje trzy niezależne diagnozy: *Formuła ABCD (klasyczna)*, *Formuła ABCD (własna optymalizacja)* oraz *Drzewo decyzji*. Na podstawie tych trzech decyzji system sugeruje wystawienie ostatecznej diagnozy wybierając wypadkową z trzech diagnoz. W przypadku, gdy wszystkie trzy diagnozy są różne, system pozostawia decyzję użytkownikowi.

4. ZAKOŃCZENIE

W pracy przedstawiono praktyczną realizację internetowego systemu diagnozowania zmian melanocytowych skóry (raka czerniaka). System ten posiadający również wyraźne funkcje dydaktyczne, oblicza wartości parametru TDS (Total Dermatoscopy Score) posługując się trzema odmiennymi algorytmami, a zatem trzema odmiennymi sposobami klasyfikacji zmian. Opracowany system pozwala na wczesną, nieinwazyjną diagnozę zmiany skórnej. Przedstawiony system jest dostępny w sieci Internet pod adresem <http://www.wsiz.rzeszow.pl/ksesi>. Dostępne są dwie wersje językowe: polska oraz angielska.

BIBLIOGRAFIA

- [1] Braun-Falco O., Stolz W., Bilek P., Merkle T., Landthaler M.: *Das Dermatoskop. Eine Vereinfachung der Auflichtmikroskopie von pigmentierten Haut-veränderungen*. Hautarzt 1990(40)131-135.
- [2] Hippe Z.S., Bajcar S., Błajdo P., Grzymała-Busse J.P., Grzymała-Busse J.W., Knap M., Paja W., Wrzesień M.: *Diagnosing Skin Melanoma: Current versus Future Directions*, TASK Quarterly 7(2003, No 2)289-293.
- [3] Michalski R.S., Bratko I., Kubat M. (Eds.): *Machine Learning and Data Mining: Methods and Applications*. J. Wiley & Sons Ltd, Chichester 1998.
- [4] Alvarez A., Bajcar S., Brown F.M., Grzymała-Busse J.W., Hippe Z.S.: *Optimization of the ABCD Formula Used for Melanoma Diagnosis* In: M.A. Kłopotek, S.T. Wierchoń, K. Trojanowski (Eds.), *Advances in Soft Computing, (Intelligent Information Processing and Web Mining)*, Springer-Verlag, Heidelberg 2003, pp. 233-240.
- [5] Bajcar S., Grzymała-Busse J.W., Grzymała-Busse W.J., Hippe Z.S.: *Diagnosis of Melanoma Based on Data Mining and ABCD Formulas*. 3rd International Conference on Hybrid Intelligent Systems (HIS'2003), Melbourne (Australia), 14-17.12.2002.
- [6] Stolz W.: *Auflichtmikroskopische Diagnose des malignen Melanoma*. W: Garbe C., Dummer R., Kaufmann R., Tilgen W. (Red.) *Dermatologische Onkologie*, Springer-Verlag, Heidelberg 1997, str. 281-304.
- [7] <http://www.dermoncology.com/dermoscopy/stolzII.html>
- [8] Quinlan J.R.: *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo (CA), 1993

INTERNET BASED SYSTEM FOR DIAGNOSIS OF MELANOMA

Summary

In this paper, the development of Internet based system for diagnosis of melanoma is shortly described. This system has distinct learning functions. He calculates values of **TDS** (Total Dermatoscopy Score) parameter using three different algorithms. This system enables users to early and non-invasion diagnosis of skin spots. This system contains a set of instructions to understand the way how to determine and/or to calculate values of TDS parameter. Discussed system is available on the Internet using URL: <http://www.wsiz.rzeszow.pl/ksesi>.

Anna Pietrenko-Dąbrowska⁽¹⁾, Renata Kalicka⁽²⁾

⁽¹⁾Katedra Systemów Mikroelektronicznych, ⁽²⁾Katedra Inżynierii Biomedycznej,
Politechnika Gdańska

OPTIMALIZACJA POBUDZEŃ DLA CELÓW IDENTYFIKACJI PARAMETRÓW KOMPARTMENTOWYCH MODELI SYSTEMÓW FARMAKOKINETYCZNYCH

Streszczenie

Pobudzenia optymalne, według kryterium optymalizacji czułościowej, wykorzystano do estymacji parametrów kompartmentowych modeli systemów farmakokinetycznych metodą minimalizacji błędów predykcji. Na pobudzenia optymalne nałożono ograniczenia na energię i czas trwania. Uzyskane, dla pobudzeń optymalnych, dokładności estymat parametrów porównano z dokładnościami uzyskanymi dla pobudzeń nieoptymalnych, standardowo stosowanych w praktyce klinicznej: pobudzenia skokowego (wlew) oraz pobudzenia typu bolus (iniekcja). Obliczenia przeprowadzono w Matlabie.

1. MODELE KOMPARTMENTOWE SYSTEMÓW FARMAKOKINETYCZNYCH

Do opisu przepływu substancji w organizmie żywym stosowane są modele kompartmentowe opisane w kategoriach zmiennych stanu. Modele te należą do klasy modeli parametrycznych. W procesie ich identyfikacji można wykorzystać posiadaną wiedzę *a priori* na temat funkcjonowania badanego systemu. Na podstawie tej wiedzy określona zostaje struktura modelu (równania różniczkowe), a następnie na podstawie danych pomiarowych estymuje się wartości liczbowe parametrów modelu.

W farmakokinytyce parametry modelu to stałe przepływu substancji pomiędzy kompartmentami oraz stałe eliminacji substancji z poszczególnych kompartmentów. Często jedynym dostępnym pomiarowym portem wejściowym i jednocześnie wyjściowym jest układ krwionośny, więc dalsze rozważania ograniczono do systemów klasy SISO (*Single Input Single Output*). Ciągły model n zmiennych stanu opisuje się układem równań stanu i równaniem wyjścia [1]

$$\dot{\mathbf{x}}(t) = \mathbf{A}(\mathbf{p}) \cdot \mathbf{x}(t) + \mathbf{B} \cdot u(t), \quad y(t) = \mathbf{C} \cdot \mathbf{x}(t) + v(t), \quad (1.1)$$

gdzie: $\mathbf{x}(t) = [x_1(t), \dots, x_n(t)]^T$ – wektor stanu o wymiarach $n \times 1$,

$u(t)$, $y(t)$, $v(t)$ – odpowiednio: pobudzenie, odpowiedź i błąd pomiaru,

- \mathbf{p} – wektor parametrów modelu o wymiarach $n_k \times 1$,
 $\mathbf{A}(\mathbf{p})$ – macierz układu o wymiarach $n \times n$,
 \mathbf{B} – macierz wejścia o wymiarach $n \times 1$,
 \mathbf{C} – macierz wyjścia o wymiarach $1 \times n$.

Dla kompartmentowych modeli farmakokinetycznych stan x_i , $i = 1, 2, \dots, n$, oznacza stężenie (masę) substancji w i -tym kompartmentcie. Wektor parametrów modelu \mathbf{p} zawiera stałe przepływu k_{ij} , $i, j = 1, 2, \dots, n$, substancji z kompartmentu j -tego do i -tego, oraz stałe eliminacji substancji z i -tego kompartmentu k_{0i} . Elementy macierzy $\mathbf{A}(\mathbf{p})$ związane są ze stałymi przepływu k_{ij} oraz k_{0i} zależnością

$$a_{ij} = k_{ij}, i \neq j; a_{ii} = -k_{0i} - \sum_{j=1}^n k_{ji}. \quad (1.2)$$

Dane pomiarowe, wykorzystywane do identyfikacji, mają postać ciągów próbek. Powoduje to, że zamiast modeli ciągłych stosuje się modele dyskretne. Model ciągły opisany równaniami (1.1) próbkowany jest ze stałym odstępem próbkowania T . Jeśli pobudzenie można uznać za stałe pomiędzy kolejnymi momentami próbkowania, wówczas dla $t = kT + T$, gdzie k oznacza kolejny numer próbki, równania stanu i wyjścia mają postać [2]

$$\mathbf{x}[kT + T] = \mathbf{F}(\mathbf{p}) \cdot \mathbf{x}[kT] + \mathbf{G} \cdot u[kT], \quad y[kT] = \mathbf{H} \cdot \mathbf{x}[kT] + v[kT], \quad (1.3)$$

gdzie

$$\mathbf{F}_{n \times n}(\mathbf{p}) = e^{\mathbf{A}(\mathbf{p})T}, \quad \mathbf{G}_{n \times 1}(\mathbf{p}) = \int_0^T e^{\mathbf{A}(\mathbf{p})(T-\tau)} \mathbf{B} d\tau, \quad \mathbf{H}_{1 \times n} = \mathbf{C}. \quad (1.4)$$

Dla modelu stochastycznego zmiennych stanu, uwzględniającego istnienie szumu procesu w oraz szumu pomiarowego v , równania stanu i wyjścia przyjmują postać [1]

$$\mathbf{x}[kT + T] = \mathbf{F}(\mathbf{p}) \cdot \mathbf{x}[kT] + \mathbf{G} \cdot u[kT] + w[kT], \quad y[kT] = \mathbf{H} \cdot \mathbf{x}[kT] + v[kT]. \quad (1.5)$$

2. PREDYKCJA

Zadanie predykcji (prognozowania jednokrokowego) polega na estymowaniu wartości odpowiedzi $y[kT + T]$ w momencie czasu $kT + T$ na podstawie ciągu próbek odpowiedzi $y_0^{kT} = \{y[-\infty], \dots, y[(k-1)T], y[kT]\}$ pobranych ze stałym odstępem próbkowania T [2], [3]. Zazwyczaj do obliczenia prognozy konieczna jest również znajomość ciągu próbek pobudzenia u_0^{kT} . W przypadku modeli zmiennych stanu zamiast estymaty odpowiedzi często poszukiwana jest estymata stanu systemu. W praktyce poszukiwana prognoza obliczana jest na podstawie skończonych ciągów próbek.

Estymatę o wariancji minimalnej (prognozę jednokrokową) $\hat{\mathbf{x}}[kT + T | kT]$ wartości, jaką przyjmie wektor stanu \mathbf{x} modelu zmiennych stanu w momencie czasu $kT + T$, można wyznaczyć na podstawie skończonego ciągu próbek $y_0^{kT} = \{y[0], y[T], \dots, y[kT]\}$ jako wartość średnią warunkową [3]

$$\hat{\mathbf{x}}[kT + T | kT] = E[\mathbf{x}[kT + T] | y_0^{kT}]. \quad (2.1)$$

Model predykcyjny (prognozę) dla modelu stochastycznego zmiennych stanu opisanego równaniami (1.5) można przedstawić w postaci rekurencyjnej [1], [3]

$$\hat{\mathbf{x}}[kT+T|kT] = \mathbf{F} \cdot \hat{\mathbf{x}}[kT|kT-T] + \mathbf{G} \cdot u[kT] + \mathbf{K} \cdot (y[kT] - \mathbf{H} \cdot \hat{\mathbf{x}}[kT|kT-T]), \quad (2.2)$$

gdzie $\mathbf{K}_{n \times 1}$ oznacza wzmocnienie filtru Kalmana

$$\mathbf{K} = \mathbf{F} \cdot \mathbf{P}[kT|kT-T] \cdot \mathbf{H}^T \cdot [\mathbf{H} \cdot \mathbf{P}[kT|kT-T] \cdot \mathbf{H}^T + R_v]^{-1}, \quad (2.3)$$

natomiast prognoza jednokrokowa wyjścia

$$\hat{y}[kT+T|kT] = \mathbf{H} \cdot \hat{\mathbf{x}}[kT+T|kT]. \quad (2.4)$$

Zależności (2.2), (2.3) i (2.4) są słuszne przy następujących założeniach: ciągi $\{\mathbf{x}[kT]\}$, $\{y[kT]\}$, $k=1, \dots, N$, mają łączny rozkład normalny, $\{u[kT]\}$ jest znanym ciągiem wejściowym, $\mathbf{x}[0]$ ma rozkład normalny $N(\mathbf{x}_0, \mathbf{P}_0)$, szum wyjścia $\{v[kT]\}$ oraz szum procesu $\{w[kT]\}$ są niezależne i dla każdego $t=kT$ mają rozkład normalny odpowiednio $N(0, R_v)$ i $N(0, R_w)$, ciąg $\{\mathbf{x}[kT]\}$ jest niezależny od $\{v[kT]\}$ oraz $\{w[kT]\}$.

Model predykcyjny (predyktor) określony zależnościami (2.2) oraz (2.4) można przedstawić w postaci innowacji (*innovations representation*) [1]

$$\hat{\mathbf{x}}[kT+T|kT] = \mathbf{F} \cdot \hat{\mathbf{x}}[kT|kT-T] + \mathbf{G} \cdot u[kT] + \mathbf{K} \cdot e[kT], \quad (2.5)$$

gdzie $e[kT]$ oznacza innowację tzn. tę część odpowiedzi, której nie można estymować na podstawie obserwacji dokonanych przed czasem kT , zatem

$$y[kT] = \mathbf{H} \cdot \hat{\mathbf{x}}[kT|kT-T] + e[kT]. \quad (2.6)$$

Definiuje się model innowacji [1]

$$\mathbf{x}[kT+T] = \mathbf{F} \cdot \mathbf{x}[kT] + \mathbf{G} \cdot u[kT] + \mathbf{K} \cdot v[kT], \quad y[kT] = \mathbf{H} \cdot \mathbf{x}[kT] + v[kT], \quad (2.7)$$

którego prognoza jednokrokowa stanu jest następująca

$$\hat{\mathbf{x}}[kT+T|kT] = \mathbf{F} \cdot \hat{\mathbf{x}}[kT|kT-T] + \mathbf{G} \cdot u[kT] + \mathbf{K} \cdot (y[kT] - \mathbf{H} \cdot \hat{\mathbf{x}}[kT|kT-T]). \quad (2.8)$$

Z powyższych zależności wynika, że dla modelu innowacji prognoza jednokrokowa spełnia warunek $\hat{\mathbf{x}}[kT+T|kT] = \mathbf{x}[kT+T]$.

3. METODA BŁĘDÓW PREDYKCJI

Zdefiniujmy parametryzowany zbiór modeli, inaczej strukturę modelu

$$\theta = \{\theta(\mathbf{p}), \mathbf{p} \in \mathbf{P}\}, \quad (3.1)$$

gdzie: $\mathbf{P} \in \mathbf{R}^n$ oznacza dopuszczalny zbiór wartości parametrów modelu, natomiast poszczególne modele należące do zbioru modeli θ oznaczone są jako $\theta(\mathbf{p})$. Model predykcyjny można zapisać jako $\theta(\mathbf{p}) = \hat{\mathbf{x}}[kT|kT-T, \mathbf{p}]$, przy czym w zapisie uwzględ-

niono zależność prognozy od wektora parametrów \mathbf{p} . Zadanie estymacji parametrów polega na znalezieniu w zbiorze modeli predykcyjnych θ pewnego modelu $\theta(\hat{\mathbf{p}})$, który zapewni jak najwierniejsze odtworzenie odpowiedzi badanego systemu na dane pobudzenie. Estymatę wektora parametrów modelu uzyskaną na podstawie N punktów pomiarowych oznacza się jako $\hat{\mathbf{p}}_N$. Błędy predykcji związane z modelem $\theta(\hat{\mathbf{p}}_N)$

$$\varepsilon[kT, \hat{\mathbf{p}}_N] = y[kT] - \hat{y}[kT | kT - T, \hat{\mathbf{p}}_N], \quad k = 1, \dots, N. \quad (3.2)$$

Metody estymacji parametrów, w których poszukiwana estymata minimalizuje skalarną miarę (normę) N -elementowego wektora błędów predykcji, nazywane są metodami minimalizacji błędów predykcji (*prediction-error methods, pem*). Norma ta stanowi miarę jakości modelu $\theta(\mathbf{p})$ i definiuje się ją ogólnie jako

$$V_N(\mathbf{p}) = \frac{1}{N} \sum_{k=1}^N l(\varepsilon[kT, \mathbf{p}]), \quad (3.3)$$

gdzie $l(\varepsilon)$ oznacza funkcjonal przyjmujący wartości dodatnie. Estymata *pem* wektora parametrów $\hat{\mathbf{p}}_N$ minimalizuje powyższą normę ogólną na zbiorze dopuszczalnych wartości parametrów \mathbf{P}

$$\hat{\mathbf{p}}_N = \arg \min_{\mathbf{p} \in \mathbf{P}} V_N(\mathbf{p}). \quad (3.4)$$

Najczęściej stosowaną funkcją $l(\varepsilon)$ jest funkcja kwadratowa [2] i [4]

$$l(\varepsilon) = \frac{1}{2} \varepsilon^2. \quad (3.5)$$

Dla tej postaci funkcji $l(\varepsilon)$ metoda minimalizacji błędów predykcji sprowadza się do metody najmniejszych kwadratów. Co więcej, jeśli (jak założono) zakłócenia pomiaru są niezależne dla każdej próbki odpowiedzi i mają rozkład normalny o średniej zero i jednakowej wariancji R , wówczas estymata najmniejszych kwadratów zgodna jest z estymatą największej wiarygodności, gdyż dla zakłóceń tego typu norma błędu (3.3) jest kwadratową funkcją błędu. Co oznacza, że w rozważanym przypadku metoda minimalizacji błędów predykcji zgodna jest zarówno z metodą najmniejszych kwadratów, jak i metodą największej wiarygodności.

4. OPTIMALIZACJA SYGNAŁU TESTUJĄCEGO

Nierówność informacyjna (twierdzenie Cramera-Rao [4]) określa możliwą do uzyskania dokładność estymat parametrów

$$\text{cov}[\hat{\mathbf{p}}] = E[(\hat{\mathbf{p}} - \mathbf{p})(\hat{\mathbf{p}} - \mathbf{p})^T] \geq \mathbf{M}^{-1}, \quad (4.1)$$

gdzie: $\mathbf{M}_{n_k \times n_k}$ to macierz informacyjna Fishera, zaś $\text{cov}[\hat{\mathbf{p}}]_{n_k \times n_k}$ oznacza kwadratową macierz kowariancji estymat parametrów, która jest jednocześnie miarą odległości pomiędzy n_k -elementowymi wektorami: wektorem estymat parametrów $\hat{\mathbf{p}}_N$ a wektorem rzeczywistych wartości parametrów \mathbf{p} .

W czasie dyskretnym elementy macierzy informacyjnej Fishera \mathbf{M} dla systemu SISO, jeśli błędy pomiarów każdej próbki odpowiedzi $y[kT]$, $k = 1, 2, \dots, N$, mają jednakowy rozkład normalny o średniej równej zero i wariancji R , opisane są zależnością

$$\mathbf{M} = [m_{ij}] = \left[\frac{1}{R} \sum_{k=1}^N \frac{\partial y[kT]}{\partial p_i} \frac{\partial y[kT]}{\partial p_j} \right], \quad i, j = 1, \dots, n_k, \quad (4.2)$$

gdzie: $\frac{\partial y[kT]}{\partial p_i}$ oznacza czułość odpowiedzi y względem parametru p_i w chwili kT .

Jak wynika z (4.1) minimalizacja macierzy kowariancji estymat parametrów równoważna jest maksymalizacji macierzy informacyjnej Fishera \mathbf{M} . W praktycznej maksymalizacji macierzy \mathbf{M} stosuje się skalarne kryteria optymalności $\Phi(\mathbf{M})$. W niniejszej pracy stosowany funkcjonal celu ma postać śladu macierzy informacyjnej Fishera. Na głównej przekątnej macierzy informacyjnej Fishera znajdują się kwadraty czułości odpowiedzi względem poszczególnych parametrów, zatem funkcjonal ten zapewnia ponadto maksymalną czułość odpowiedzi względem estymowanych parametrów. Poszukujemy pobudzenia u , które zapewni maksimum kryterium $\Phi(\mathbf{M})$ i jednocześnie, z przyczyn praktycznych, ograniczony jest czas jego trwania $T_u = N \cdot T$, oraz energia E . Funkcjonał celu w optymalizacji czułościowej oraz ograniczenia dla T_u i E mają postać

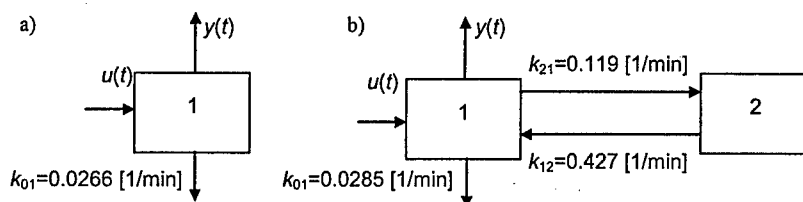
$$\Phi(\mathbf{M}) = \text{trace} \mathbf{M} = \sum_{i=1}^{n_k} m_{ii} = \frac{1}{R} \sum_{i=1}^{n_k} \sum_{k=1}^N \left(\frac{\partial y[kT]}{\partial p_i} \right)^2, \quad (4.3)$$

$$T_u = \text{const}, \quad E = T \sum_{k=1}^N u^2[kT] = \text{const}. \quad (4.4)$$

Funkcjonał celu (4.3) oraz ograniczenia (4.4) są nieliniowe, zatem zagadnienie optymalizacji sygnału testującego stanowi zagadnieniem programowania nieliniowego z ograniczeniami (*Nonlinear Programming*; NLP). Zadanie rozwiązuje się metodami wykorzystującymi warunki konieczne Kuhna-Tuckera będące warunkami koniecznymi i wystarczającymi optymalności wypukłego problemu NLP. Wykorzystywane są one przez wybraną do obliczeń procedurę *fmincon* Matlaba.

5. IMPLEMENTACJA

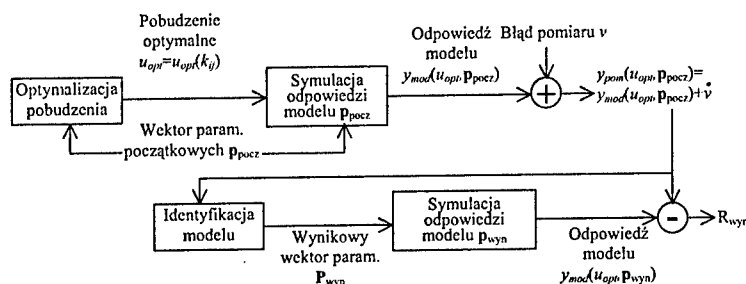
Badano 1- i 2-kompartментowy model dystrybucji gonadotropiny jak na rys.1.



Rys. 1. Badane modele gonadotropiny 1-kompartментowy (a) i 2-kompartментowy (b).

Początkowe wektory parametrów \mathbf{p}_{pocz} obu modeli zamieszczone na rys. 1 estymowano na podstawie eksperymentu intuicyjnego, nieoptymalnego [5]: podano pobudzenie w postaci iniekcji o dawce $D = 25$ i pobrano 23 próbki krwi w czasie 117 h. Wartości parametrów estymowano poprzez minimalizację sumy kwadratów odchylek pomiędzy

pomiarami oraz funkcją opisującą odpowiedź modelu [6]. Zaimplementowana procedura reestymacji wektora parametrów modeli 1- i 2-kompartmentowego została przedstawiona na rys. 2.

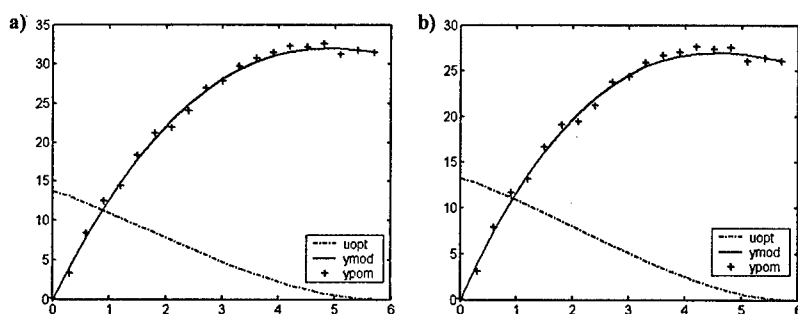


Rys. 2. Zaimplementowana procedura reestymacji wektora parametrów.

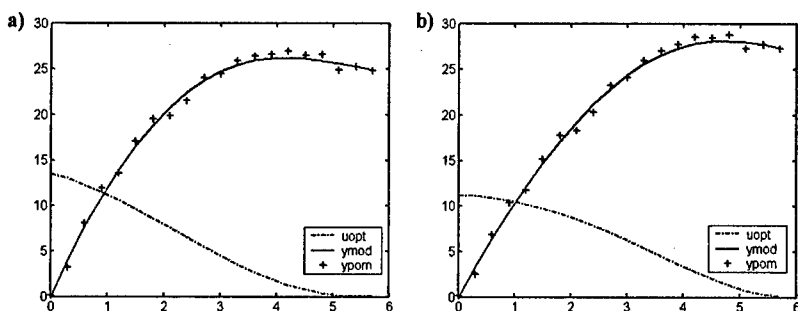
Najpierw na podstawie początkowego wektora parametrów \mathbf{p}_{pocz} , wyznaczono, przy wykorzystaniu procedury *fmincon* Matlaba, pobudzenia optymalne u_{opt} według kryterium optymalizacji czułościowej dla wybranego modelu i wybranej stałej przepływu. Dla modelu 1-kompartmentowego obliczono pobudzenie optymalne dla estymacji stałej k_{01} , natomiast dla modelu 2-kompartmentowego obliczono trzy różne pobudzenia optymalne dla estymacji poszczególnych stałych k_{01} , k_{12} i k_{21} . Przyjęto ograniczenia: czas trwania pobudzenia $T_u = 6$ min i energię $E = 300$. Energia ta odpowiada energii pobudzenia zastosowanego w eksperymencie intuicyjnym, w którym dawkę leku 25 podano w iniekcji trwającej 2 minuty. Zasymulowano odpowiedź modelu o wektorze parametrów \mathbf{p}_{pocz} na pobudzenie optymalne u_{opt} oraz na pobudzenia nieoptymalne, standardowo stosowane w farmakokinytyce: uskok (infuzja) u_{step} o czasie trwania T_u oraz bolus u_{bolus} (iniekcję) o czasie trwania T . Wszystkie rozważane pobudzenia spełniają ograniczenie $E = const$ (*equienergy inputs*). Do odpowiedzi na powyższe pobudzenia dodano szum v o rozkładzie $N(0, R_{pocz})$, gdzie $R_{pocz} = 0.3$ i jest zbliżona do wariancji uzyskanej w eksperymencie intuicyjnym. Otrzymano zaszumione odpowiedzi $y_{pom}(u, \mathbf{p}_{pocz})$, gdzie u to kolejno u_{opt} , u_{step} i u_{bolus} . Na podstawie próbek pobudzenia i odpowiedzi przeprowadzono identyfikację parametrów modeli 1- i 2-kompartmentowego przedstawionych na Rys. 1. metodą minimalizacji błędów predykcji. Obliczenia przeprowadzone zostały w Matlabie przy wykorzystaniu procedury *pem*. Estymowano parametry modeli w postaci innowacji opisanej zależnością (2.7) przy założeniu zerowego szumu procesu. Dla modelu 1-kompartmentowego wektor parametrów wynikowych \mathbf{p}_{wyn} zawiera reestymowaną stałą k_{01} . Natomiast dla modelu 2-kompartmentowego \mathbf{p}_{wyn} zawiera 3 parametry k_{01} , k_{12} i k_{21} . Kolejno każdy z nich jest reestymowany, a pozostałe mają wartości początkowe.

6. WYNIKI

Pobudzenia optymalne u_{opt} względem k_{01} dla modeli 1- i 2-kompartmentowego oraz względem k_{12} i k_{21} dla modelu 2-kompartmentowego przedstawione są odpowiednio na rys. 3. i rys. 4.



Rys. 3. Pobudzenie optymalne u_{opt} względem k_{01} , $E = 300$, $T_u = 6 \text{ min}$, $y_{pom}(u_{opt}, p_{pocz})$ oraz $y_{mod}(u_{opt}, p_{wyn})$ dla modelu 1-kompartmentowego (a) oraz 2-kompartmentowego (b).



Rys. 4. Pobudzenie optymalne u_{opt} względem k_{12} (a) oraz k_{21} (b), $E = 300$, $T_u = 6 \text{ min}$, $y_{pom}(u_{opt}, p_{pocz})$ oraz $y_{mod}(u_{opt}, p_{wyn})$ dla modelu 2-kompartmentowego.

Szczegółowe wyniki estymacji zamieszczono w tablicy 5.1. Tablica ta zawiera wartości estymat parametrów, ich wariancje oraz współczynniki zmienności (współczynniki rozproszenia zmiennej). Wariancje parametrów estymowano jako odwrotność macierzy informacyjnej Fishera (zależności (4.1) i (4.2))

$$\text{var}(k_{ij}) = \left[\frac{1}{R_{wyn}} \sum_{k=1}^N \left(\frac{\partial y_{mod}[kT]}{\partial k_{ij}} \right)^2 \right]^{-1}, \quad R_{wyn} = \frac{1}{N} \sum_{k=1}^N (y_{pom}[kT] - y_{wyn}[kT])^2, \quad (5.1)$$

gdzie R_{wyn} oznacza wariancję residuów w rozwiązaniu. Natomiast współczynnik zmienności zdefiniowany jest jako

$$CV k_{ij} [\%] = \frac{\sqrt{\text{var } k_{ij}}}{k_{ij \text{ pocz}}} \cdot 100\%. \quad (6.1)$$

Zamieszczone w ostatnim wierszu tablicy 5.1 wyniki dla bolusa odpowiadają eksperymentowi intuicyjnemu, w którym dawkę leku 25 podano w iniekcji trwającej 2 minuty. Na uzyskane dokładności estymat parametrów wpływ ma przyjęta w obliczeniach wariancja szumu symulującego błąd pomiaru odpowiedzi, która jest zbliżona do wariancji uzyskanej w eksperymencie intuicyjnym.

Tablica 5.1

Estymaty parametrów modelu 1- i 2-kompartimentowego oraz ich dokładności

Model	1-kompart.		2-kompartimentowy									
	k_{01}	$\text{var } k_{01} \cdot 10^{-6}$	$CV[\%]$	k_{01}	$\text{var } k_{01} \cdot 10^{-6}$	$CV[\%]$	k_{12}	$\text{var } k_{12} \cdot 10^{-3}$	$CV[\%]$	k_{21}	$\text{var } k_{21} \cdot 10^{-6}$	$CV[\%]$
u_{opt}	0.0249	3.99	7.51	0.0266	6.05	8.57	0.454	1.07	7.65	0.1144	22.67	4.00
u_{step}	0.0246	6.51	9.59	0.0262	10.40	11.23	0.460	2.04	10.57	0.1139	34.36	4.93
u_{bolus}	0.0246	4.83	8.26	0.0262	8.10	9.92	0.460	1.42	8.83	0.1134	31.91	4.75

Wyniki pokazują, że dla modelu 1- i 2-kompartimentowego oraz dla każdego parametru, pobudzenie optymalne zapewnia najmniejsze wartości wariancji oraz współczynnika zmienności estymat parametrów niż standardowo stosowane w praktyce klinicznej pobudzenie typu bolus (iniekcja) oraz skokowe (infuzja) o tej samej energii. Dla niektórych zastosowań dokładność uzyskiwana w przypadku pobudzeń standardowych może okazać się wystarczająca. Jeśli jednak istotne jest, aby parametry modelu były estymowane z możliwie największą dokładnością, wówczas do identyfikacji modelu należy zastosować pobudzenie optymalne.

BIBLIOGRAFIA

- [1] L. Ljung: System identification – Theory for the user, Prentice Hall, 1999,
- [2] P. Eykhoff: Identyfikacja w układach dynamicznych, PWN, 1980,
- [3] B.D.O. Anderson, J.B. Moore: Filtracja optymalna, WNT, 1984,
- [4] S. Brandt: Analiza danych, PWN, 1999,
- [5] J.E.A. McIntosh, R.P. McIntosh: Mathematical Modelling and Computers in Endocrinology, Springer-Verlag, 1980,
- [6] R. Kalicka: Modelowanie procesów kinetycznych w systemach biomedycznych, Wydawnictwo Politechniki Gdańskiej, 2000.

THE COMPARISON OF OPTIMAL INPUTS FOR PARAMETER ESTIMATION OF COMPARTMENTAL MODELS OF PHARMACOKINETIC SYSTEMS

Summary

Optimal input design using sensitivity criterion for parameter estimation of compartmental models of pharmacokinetic systems is presented. The energy of the signal and the signal duration are constrained. The prediction error method was used to estimate parameters' values for 1- and 2-compartmental models. The achieved results are compared to results obtained for non-optimal routine inputs: step input and bolus input. The calculations were performed using Matlab.

Jacek Rumiński, Barbara Bobek-Billewicz*

**Katedra Inżynierii Biomedycznej, Politechnika Gdańska,
(*) Samodzielna Pracownia Neuroradiologii, Akademia Medyczna w Gdańsku**

SYNTEZA OBRAZÓW PARAMETRYCZNYCH W BADANIU PERFUZJI MÓZGU METODĄ MRI

Streszczenie

Zaprezentowano analizę teoretyczną podstaw obrazowania parametrycznego dla badań dynamicznych mózgu z dożylnym podaniem środka kontrastującego. Opracowano oprogramowanie umożliwiające syntezę obrazów parametrycznych. Na podstawie przeprowadzonej analizy oraz wstępnych wyników symulacji wskazano na potrzebę dalszych badań celem standaryzacji obrazowania parametrycznego.

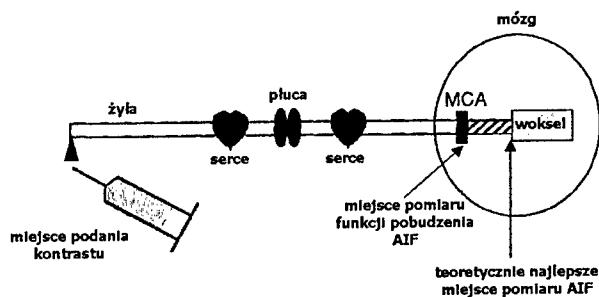
1. WSTĘP

Istniejące metody diagnostyki ośrodkowego układu nerwowego (OUN) bazują na jednostkowych pomiarach medycznych poprzez rozproszone systemy pomiarowe. Powszechnie stosowane są w ramach diagnostyki OUN sformalizowane klasyfikacje i oceny [1] uzupełniane przez dodatkowe badania neuroobrazowe (głównie CT lub MRI) [2]. Bardzo istotną, a często kluczową metodą badania mózgu jest badanie przepływów i metabolizmu mózgowego (TCD, SPECT, PET) [3]. Wiele źródeł wskazuje równocześnie iż uwaga powinna być skupiona głównie na badaniach PET oraz MRI. W sposób szczególny podkreślona jest rola obrazowania parametrycznego (np. CBV – *Cerebral Blood Volume*, CBF – *Cerebral Blood Flow*, MTT – *Mean Transit Time*) jak i techniki spektroskopii MR oraz konieczność poszukiwania nowych radiofarmaceutyków. W literaturze [4] wyraźnie stwierdza się konieczność stosowania metod PET i MRI/fMRI do diagnostyki chorób mózgu, głównie demencji. Podkreśla się także [5] rolę wspólnej analizy PET i MRI/CT dla diagnostyki mózgu.

W badaniach przepływów mózgowych i ocenie dynamiki zdefiniowano szereg klas obrazów parametrycznych jak CBF, CBV, MTT powstałych przez opis krzywych zmian aktywności kontrastu w tkankach dla analizowanej serii czasowej obrazów tego samego obiektu [6,7]. Prezentowana praca skupia się na opisie i syntezie obrazów parametrycznych w technice MRI.

2. DYNAMICZNY POMIAR MRI

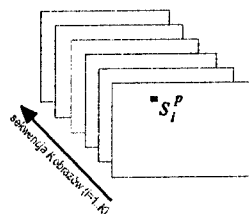
W wyniku podania kontrastu zmieniają się własności wielkości mierzonej (zmiana właściwości magnetycznych ośrodka) dla obserwowanego obiektu. Kontrast, podawany dożylnie (rys. 1), np. Gd-DTPA, transportowany jest z krwią do mózgu, gdzie ulega dalszej dystrybucji. W wyniku pomiaru rejestrowana jest różnica sygnałów pochodzących od molekuł środka kontrastowego i tkanek, zgromadzonych w mierzonych kompartmentach (wokselach \rightarrow pikselach). Wprowadzone, przez odmienne właściwości magnetyczne podanego środka, zmiany są widoczne na obrazie. Większa jest zatem również różnica (gradient, kontrast) pomiędzy tkankami zawierającymi, lub nie, podany znacznik. Taka metoda pomiarowa nazywana jest MRI Dynamic Susceptibility Contrast (dynamiczny kontrast podatności magnetycznej), w skrócie MRI-DSC.



Rys. 1. Podanie kontrastu o odmiennych właściwościach magnetycznych.
Opis w tekście.

Inną grupą metod podawania środka kontrastowego jest taka modyfikacja transportowanych do tkanek molekuł, aby można było kontrastować ich rozkład w tkankach. Do najbardziej znanych metod tej grupy należą metody BOLD (*Blood Oxygen Level Dependent*) i ASL/AST (*Arterial Spin Labelling/Tagging*).

W wyniku pomiaru w MRI-DSC (stosując zwykle szybkie techniki skanowania, głównie SE-EPI i GRE-EPI, obrazy $T2^*$ zależne – w opisywanych badaniach stosowano pomiar DSC-MRI 1.5T SE-EPI, TR=1430ms, TE=43ms, 12 warstw, 60 obrazów w serii), uzyskujemy serię obrazów. Ponieważ przejście kontrastu Gd-DTPA przez woksel trwa około 60 sekund, stąd okres ten jest próbkowany określonym zestawem regularnie rejestrowanych obrazów ($60 \cdot 1s$). Rozdzielczość przestrzenna zależy od możliwości systemu pomiarowego. Zwykle wynosi 128×128 pikseli (może być wyższa kosztem rozdzielczości czasowej-liczby obrazów w okresie badania). Załóżmy, że rozpatrujemy pojedynczy punkt (piksel) obiektu. Wówczas możliwa jest reprezentacja zmian wartości mierzonego punktu w czasie (rys.2).



Rys. 2. Sekwencja pomiarowa: S – wartość pomiarowa dla i -tej warstwy dla punktu p .

Uzyskana funkcja reprezentuje dynamikę zachodzącego zjawiska w badanym punkcie i jest określana jako TAC (*Time Attenuation Curves*) lub TTAC (*Tissue Time Activation Curve*). Polaryzacja krzywej TAC zależy od stosowanego kontrastu i metody pomiarowej. Dla Gd-DTPA i pomiaru T2 polaryzacja jest ujemna (tłumienie rejestrowanego sygnału). Podstawową operacją przed przystąpieniem do obliczeń obrazów parametrycznych jest określenie krzywej koncentracji środka kontrastującego. Na podstawie badań przyjęto, że natężenie sygnału wywołane podaniem środka kontrastowego dane jest zależnością

$$\begin{aligned} S_K &= \rho \exp\left(-\frac{TE}{T2_K^*}\right) = \rho \exp\left(-TE \cdot \left(\frac{1}{T2_0^*} + \frac{1}{T2_{K+}^*}\right)\right) = \\ &= \rho \exp\left(-\frac{TE}{T2_0^*}\right) \cdot \exp\left(-\frac{TE}{T2_{K+}^*}\right) = \rho \exp\left(-\frac{TE}{T2_0^*}\right) \cdot \exp(-TE \cdot k \cdot C_k) = \\ &= S_0 \cdot \exp(-TE \cdot k \cdot C_k), \end{aligned} \quad (2.1)$$

gdzie: ρ – gęstość cząstek w badanym obiekcie posiadających spin; $T2_K^*$ czas $T2^*$ dla pomiaru obiektu ze środkiem kontrastowym; $T2_{K+}^*$ teoretyczny czas $T2^*$ stanowiący wartość dodaną w wyniku podania (obecności) środka kontrastującego, odwrotnie proporcjonalny do koncentracji kontrastu C_k ; k – współczynnik proporcjonalności, S_0 – sygnał $T2^*$ zależny rejestrowany dla danego obiektu bez obecności środka kontrastowego.

Wielkość $\left(\frac{1}{T2_0^*} + \frac{1}{T2_{K+}^*}\right)$, oznacza się także jako $\Delta R2^*$ ($R=1/T$).

Zatem na podstawie powyższych wzorów łatwo można wyznaczyć funkcję opisującą zmianę koncentracji, bazując na pomiarach z i bez podania kontrastu:

$$C_k(t) = -\frac{1}{k \cdot TE} \ln\left(\frac{S_k(t)}{S_0}\right) = -\frac{1}{k \cdot TE} \ln(SA(t)), \quad (2.2)$$

gdzie SA – (*Signal Attenuation*) tłumienie sygnału.

Współczynnik proporcjonalności k zależy od właściwości ośrodka i warunków pomiarowych (np. natężenia pola, parametrów sekwencji pomiarowej). Przyjmowany jako stały dla wszystkich badanych wokseli wprowadza trudny do oszacowania błąd, powodując niestety trudność określenia dokładnych wartości parametrów.

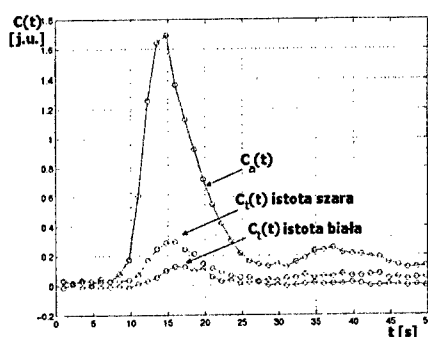
3. OPIS FUNKCJI KONCENTRACJI KONTRASTU – METODY SYNTEZY OBRAZÓW PARAMETRYCZNYCH

W rozważaniach nad perfuzją MRI dokonuje się podziału na małe naczynia (naczynia włosowate – mózg) i duże naczynia (np. MCA – *Middle Cerebral Artery*). Z tego powodu wyznacza się dwie funkcje koncentracji odpowiednio:

$$\text{– koncentracja znacznika w naczyniach tkanki: } C_t(t) = -\frac{1}{k_t \cdot TE} \ln\left(\frac{S_t(t)}{S_{t0}}\right), \quad (3.1)$$

– koncentracja znacznika w tętnicy $C_a(t) = -\frac{1}{k_a \cdot TE} \ln \left(\frac{S_a(t)}{S_{a0}} \right)$, (3.2)

gdzie k_a i k_t to współczynniki proporcjonalności omówione wyżej. Przykładowe krzywe koncentracji pokazano na rysunku 3.



Rys. 3. Przykładowe krzywe koncentracji $C_t(t)$, $C_a(t)$.

Synteza obrazów parametrycznych bazuje na budowie krzywej $C_t(t)$. Istnieją różne modele określające opis i sposób wyznaczania parametrów perfuzji tkanek mózgu. W tej pracy rozpatrzmy DSC-MRI. Załóżmy, że rozpatrywana objętość (VOI – *volume of interest*) składa się z fragmentu tkanki (sieci naczyń), do którego doprowadzone jest pobudzenie (iniekcja, wejście układu) i z którego wyprowadzona jest odpowiedź (wyjście układu). Zakładając, że dana jest odpowiedź impulsowa badanego obiektu (w uproszczeniu obiektem jest zestaw naczyń – rur), która w rozpatrywanej aplikacji opisuje transport masy (znacznik) przez obiekt (opóźnienia i dyspersja):

$$\int_{t=-\infty}^{\infty} h(t) \cdot dt = 1, \quad (3.3)$$

wówczas średni czas przejścia znacznika przez daną objętość można wyznaczyć jako wartość średnią dla rozkładu, czyli

$$MTT = \frac{\int_{t=0}^{\infty} h(t) \cdot t \cdot dt}{\int_{t=0}^{\infty} h(t) \cdot dt} = \int_{t=0}^{\infty} h(t) \cdot t \cdot dt. \quad (3.4)$$

Część znacznika pozostającego w badanej objętości po czasie t od podania kontrastu na wejście układu opisywana jest ilościowo przez funkcję (*residue function*):

$$R(t) = 1 - H(t) = 1 - \int_0^t h(\tau) \cdot d\tau, \quad (3.5)$$

gdzie $H(t)$ jest funkcją opisującą ile znacznika opuściło daną objętość po czasie t od podania kontrastu. Skoro $H(t)$ przyjmuje wartości z zakresu od 0 do 1, zatem $R(t)$ przyjmie wartości od 1 do 0 (w jednej chwili czasu $\tau = 0$, cały kontrast znajduje się w VOI, a potem obserwujemy tylko jego wydalenie). Ponieważ stosowane praktycznie pobudzenie nie jest

idealnym impulsem oraz odpowiedź układu mierzona jest w obrębie VOI, dlatego koncentrację kontrastu w tkance po czasie t od jego podania opisuje operacja splotu:

$$C_t(t) = \int_0^t C_a(\tau) \cdot R(t-\tau) d\tau, \quad (3.6)$$

aproxymowana przez:

$$C_t(t_n) = \left(\sum_{m=0}^n C_a(t_m) \cdot R(t_n - t_m) \right) \cdot \Delta t = \left(\sum_{m=0}^n C_a(m \cdot \Delta t) \cdot R(\Delta t \cdot (n-m)) \right) \cdot \Delta t = \left(\sum_{m=0}^n C_a[m] \cdot R[n-m] \right) \cdot \Delta t \quad (3.7)$$

gdzie $n = 0 \dots N-1$; N – liczba próbek sygnału (liczba obrazów zarejestrowanych w czasie dla danej warstwy i danego elementu objętościowego).

Powyższą aproksymację możemy zapisać równaniem macierzowym

$$C_t = C_a \times R, \quad (3.8)$$

czyli teoretycznie jego rozwiązanie (dane: C_n , C_a ; nieznane $R(t)$) można uzyskać poprzez operację:

$$R = C^{-1}_a \times C_t. \quad (3.9)$$

Jeżeli macierz C_a ($N \times N$) nie jest osobiwa (matematycznie i numerycznie) wówczas rozwiązanie tego równania dla danych C_a i C_t da wartości R . Dla rzeczywistego eksperymentu z DSC-MRI okaże się, że wartości $R(t)$ nie mieszczą się w zakresie 1-0. Dla $t = 0$ wartość R nie jest równa 1, tylko powiedzmy F , modyfikując opis matematyczny na:

$$C_t(t) = \int_0^t C_a(\tau) \cdot (F \cdot R(t-\tau)) d\tau, \quad (3.10)$$

lub w postaci aproxymowanej

$$C_t(t_n) = \left(\sum_{m=0}^n C_a[m] \cdot F \cdot R[n-m] \right) \cdot \Delta t. \quad (3.11)$$

Ponieważ F nie zależy od czasu, jest zatem współczynnikiem skali, zależnym od właściwości danego elementu objętości. Wartość F jest interpretowana jako wartość CBF, a jej wyznaczenie dane jest relacją:

$$F \cdot R(t=0) = F = CBF. \quad (3.12)$$

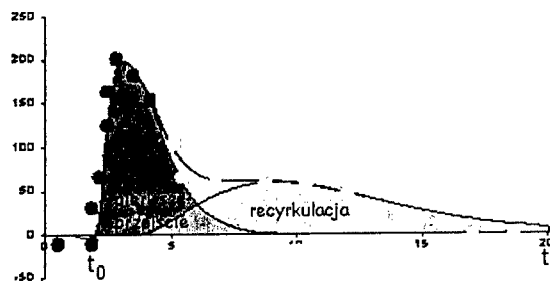
W zależności od wartości macierzy C_a obliczane w celu odwracania macierzy $C_a[m]$ wyznaczniki mogą przyjmować wartość zero (prowadzić to może to dzielenia przez 0). Ten sam rezultat można uzyskać na drodze numerycznej (skończona dokładność liczb rzeczywistych w komputerze) ze względu na bardzo małe wartości licznika. Tradycyjne próby rozwiązania tego problemu prowadzą do regularyzacji macierzy C_a , w celu eliminacji osobiwości (np. SVD). Dokładność $F \cdot R(t)$ uzyskiwanej poprzez zastosowanie transformacji Fouriera (zazwyczaj poprzez zastosowanie szybkich algorytmów FFT) silnie zależy od poziomu szumu w sygnałach $C_a(t)$ i $C_t(t)$. Stąd często lepsze wyniki daje metoda SVD. Inne metody uzyskania $F \cdot R(t)$ to próby odgadnięcia rodzaju funkcji $R(t)$ (np. $R(t) = F \cdot \exp(-t/MTT)$; parametr – MTT) poprzez testowanie wielu kandydatów i dopasowywanie

parametrów funkcji kandydujących do znanej zależności na $C_f(t)$. Metody takie nazywane są metodami parametrycznymi.

Warto tu jednak zauważyć, że mierzona w technice MRI-DSC wartość $C(t)$ wskazuje nie tylko na koncentrację kontrastu w pierwszym jego przejściu przez element objętości ale również zawiera informacji o recyrkulacji znacznika. Jak wspomniano w założeniach metody konieczna jest eliminacja udziału recyrkulacji w wartości funkcji $C(t)$. W tym celu wykorzystuje się metodę poszukiwania prawdziwej wartości $C(t)$ poprzez jej zamodelowanie, a następnie poszukiwanie takich parametrów modelu, które najlepiej (najdokładniej – najmniejszy błąd) dopasują założony model do danych pomiarowych $C(t)$, np. $C_f(t)$. Najbardziej popularną funkcją modelową $C(t)$ jest funkcja:

$$C(t) = \begin{cases} K(t-t_0)^\beta \cdot e^{-\alpha(t-t_0)}, & t > 0 \\ 0, & t < 0 \end{cases}, \quad (3.13)$$

gdzie K , α , β to parametry modelu (np. $\beta = 3$, $\alpha = 2/3$); t_0 - czas wzrostu funkcji $C(t)$ jak to pokazano na rysunku 4.



Rys. 4. Krzywe $C(t)$ z wskazaniem teoretycznej funkcji pierwszego przejścia i recyrkulacji.

Uzyskanie $R(t)$ wymaga określenia $C_d(t)$, zwanej funkcją pobudzenia, czy tętniczą funkcją wejścia AIF. Określenie AIF jest trudne. Często jako $C_d(t)$ przyjmuje się zmierzoną wartość $C(t)$ dla MCA, a właściwie blisko MCA (rys. 1). Dlaczego? Po pierwsze dlatego, że ruch molekuł w dużym naczyniu jakim jest MCA daje mierzalny sygnał NMR, zatem pomiar koncentracji w naczyniu dałby w rezultacie nadmiarowy wynik $C_d(t)$. Istnieją jeszcze inne problemy związane z pomiarem $C_d(t)$ w dużych naczyniach: zmiana mierzonego sygnału $C(t)$ w zależności od położenia naczynia względem pola, zmiana mierzonego sygnału $C(t)$ w zależności od średnicy naczynia, itp. Stąd ilościowy opis perfuzji w MRI-DSC jest bardzo trudny. W celu estymacji opisu ilościowego (w danych jednostkach) wprowadzono współczynnik skalowania $C_f(t)$ jako:

$$C_f(t) = \frac{\rho}{Kh} \int_0^t C_a(\tau) \cdot (F \cdot R(t-\tau)) d\tau, \quad (3.14)$$

gdzie: ρ – średnia gęstość tkanki mózgu $\rho = 1,04$ g/mol; Kh – relacja hematokrytów dużych i małych naczyń, $Kh = (1 - Hd)/(1 - Hm)$; $Hd = 0,45$; $Hm = 0,25$.

W ten sposób uzyskuje się mianowane wartości CBF w ml/100g/min; a CBV w ml/100g. Czym jest CBV (*Cerebral Blood Volume*)? CBF określa ilość krwi przemieszczającej się przez daną objętość tkanki w czasie. CBV natomiast oznacza całkowitą ilość krwi w danej objętości tkanki. Zatem CBV można wyznaczyć jako:

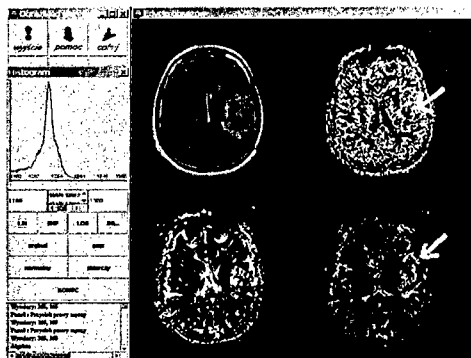
$$CBV = \frac{\int_0^{\infty} C_i(\tau) \cdot d\tau}{\frac{\rho}{Kh} \int_0^{\infty} C_a(\tau) \cdot d\tau}, \quad (3.15)$$

uwzględniając oczywiście jedynie pierwsze przejście kontrastu przez element objętości w $C(t)$. Wartość parametru MTT (*Mean Transit Time* – średni czas przejścia) można wyznaczyć na podstawie twierdzenia o centralnej objętości (*central volume theorem*)

$$MTT = CBV / CBF. \quad (3.16)$$

4. IMPLEMENTACJA METOD

Przygotowano w środowisku Java oprogramowanie syntezy i wizualizacji obrazów parametrycznych. Zaimplementowano metody generacji obrazów parametrów CBF oraz CBV (rys.5). Przygotowano również oprogramowanie dopasowania danych pomiarowych do funkcji (n -eksponencjalnej). Ze względu na szybką zbieżność oraz stabilność wybrano metodę Marquardta wraz z jej modyfikacją zaproponowaną przez Bevingtona [8]. Na podstawie wybranej metody opracowano algorytm dopasowania i zaimplementowano go w środowisku Java (Sun, JDK 1.4.1).



Rys. 5. Przykładowe obrazy MRI – strukturalny oraz obrazy CBF i CBV.

Analiza map parametrów (obrazy kolorowe!) dostarcza istotnych informacji na temat charakteru i rozległości zmian chorobowych, co ułatwia postawienie właściwej diagnozy. Przy czym perfuzja mózgowa jest tylko jednym z typów danych parametrycznych uzyskiwanych w czasie badania MR.

5. ZAKOŃCZENIE

Wciąż brak jest wspólnego standardu zarówno w zakresie definicji obrazów parametrycznych, ich doboru w procesie diagnostycznym, i stosowania. Nawet tak podstawowe obrazy parametryczne jak CBF, są często różnie tworzone w zależności od dostawcy oprogramowania. Niektórzy autorzy wskazują na konieczność poszukiwania optymalnych

metod dopasowania modelu do danych i właściwego doboru funkcji kryterialnych [9] [10]. Inni rozpatrują problem optymalnej filtracji obrazów parametrycznych w celu uzyskania możliwie jak największej rozdzielczości obrazów przy jednoczesnym dużym stopniu powtarzalności otrzymywanych map [11]. Mamy nadzieję, że prowadzone dalsze prace w tym zakresie (szczególnie dla danych MRI i PET) pozwolą nam wypracować nowe standardy diagnostyczne w zakresie obrazowania parametrycznego w badaniach dynamicznych mózgu.

BIBLIOGRAFIA

- [1] Bamford J., Sandercock P., Dennis M., Burn J., Warlow C.: Classification and natural history of clinically identifiable subtypes of cerebral infarction. *Lancet* 1991, 337, 1521-1526.
- [2] Członkowska A.: Postępy w medycynie w 1997 roku. *Neurologia Med. Prakt.* 1997, 12, 73-77.
- [3] DeWitt L.D.: Transcranial Doppler, *Stroke* 1988, 19, 915-921.
- [4] Petrella J.R., fMRI Tracks Signs of Early Dementia, *RSNA News*, <http://www.rsna.org/publications/rsnanews/sept02/fmri-1.html>, 2002.
- [5] Przetak Ch., Baum R.P., Slomka P.J., Image fusion raises clinical value of PET, *Diagnostic imaging Europe*, vol. 17 No 5, str. 10-15, 2001.
- [6] Martel A.L., Moody A.R., Allder S.J., Delay G.S., Morgan P.S., Extracting parametric images from dynamic contrast-enhanced MRI studies of the brain using factor analysis, *Medical Image Analysis*, 5:29-39, 2001.
- [7] Smith A.M., Grandin C.B., Duprez T., Mantaigne F., Cosnard G., Whole brain quantitative CBF and CBV measurements using MRI bolus tracking: Comparison of methodologies, *Magn Res Med* 43(4) pp. 559-564, 2000.
- [8] Bevington P.R., Robinson D.K.: *Data Reduction and Error Analysis for The Physical Sciences*, McGraw-Hill Higher Education, 1991.
- [9] Zhou Y., Huang S-C., Bergsneider M., Model fitting with spatial constraint for parametric imaging in dynamic PET studies, *Proc. of Brain99 Conference*, Abstract No 760, Copenhagen 1999.
- [10] Zhou Y., Huang S-C., Bergsneider M., Wong D.F., Model fitting with spatial constraint for parametric imaging in dynamic PET studies, *NeuroImage* 15, pp. 697-707, 2002.
- [11] Worsley K.J., Marrett S., Neelin P., Evans A.C., Searching Scale Space for Activation in PET images, *Human Brain Mapping* 4:74-90, 1996.

SYNTHESIS OF PARAMETRIC IMAGES IN BRAIN PERFUSION STUDIES USING MRI

Summary

Theoretical analysis of contrast-based parametric imaging for dynamic brain studies is presented. Parametric image synthesis software has been developed. Based on theoretical analysis and preliminary research results it is concluded that the further studies are required toward standardization of parametric imaging.

Patryk Babiarz, Jacek Jakiela, Maciej Piotrowski, Bartosz Pomianek

Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie

WEB SERVICES FRAMEWORK – STANDARDY, KORZYŚCI IMPLEMENTACJI ORAZ PRZYKŁADY ZASTOSOWAŃ

Streszczenie

Na przestrzeni ostatnich lat, kluczowym determinantem efektywnego funkcjonowania przedsiębiorstw stała się integracja oprogramowania w obrębie danej firmy oraz na zewnątrz z oprogramowaniem stosowanym przez partnerów biznesowych. Rozwijane przez niezależne firmy metody integracji różnych systemów, skazane są w dłuższej perspektywie na porażkę ze względu na fakt, iż wymagają dużych umiejętności programistycznych oraz są z reguły pracochłonne.

Rozwiązaniem tego problemu jest wykorzystanie w trakcie integracji tzw. usług sieciowych (ang. *Web Services*). Celem tego artykułu jest omówienie istoty usług sieciowych pod kątem stosowanych standardów, obszarów zastosowań oraz potencjalnych korzyści ich wdrożenia. W dalszej części artykułu scharakteryzowano podstawowe narzędzia wspierające implementację z wykorzystaniem *Web Services* oraz podano praktyczne przykłady zastosowań tego typu rozwiązań.

1. WSTĘP

Relatywnie prostym oraz uniwersalnym rozwiązaniem problemów związanych z integracją oprogramowania jest zastosowanie tzw. usług sieciowych (ang. *Web Services Framework*), które umożliwiają aplikacjom płynną współpracę w heterogenicznym środowisku. Usługi sieciowe można zdefiniować jako standardowy interfejs, który pozwala jednej aplikacji programowo odkryć, zinterpretować i wykorzystać usługi oferowane przez inne platformy aplikacyjne, w sposób niezależny od języka programowania, w jakim zostały stworzone [1].

Rosnącą rolę usług sieciowych podkreśla zdanie kierownika firmy Cape Clear Software, jednego z głównych dostawców rozwiązań do tworzenia usług sieciowych – „Analizując pojawienie się technologii *Web Services Framework* oraz rozpatrując jej szczegóły nie należy zapominać o jej istocie. Kluczowym aspektem usług sieciowych jest uniwersalny oraz powszechny język dla integracji oprogramowania, który zainicjował ogromne zmiany w sposobie, w jaki aplikacje komunikują się wzajemnie poprzez sieć.”

Pomimo faktu, iż usługi sieciowe oparte na bazie języka XML (ang. *eXtensible Markup Language*) są stosunkowo proste w implementacji, a ich zastosowanie może przynieść wiele korzyści, doszło do paradoksalnej sytuacji. Na rynku pojawiło się bardzo dużo narzędzi wspomagających integrację systemów w oparciu o usługi sieciowe (m.in. pro-

dukty firm IBM, Microsoft, Oracle, Sun Microsystems), natomiast zainteresowanie nimi jest wciąż marginalne. Wiąże się to z faktem, iż zagadnienia związane z usługami sieciowymi wciąż nie są powszechnie znane pracownikom działów IT. W związku z tym artykuł ten ma charakter głównie przeglądowy, a jego głównym celem jest zwrócenie uwagi na zmiany zachodzące w sposobie integracji oprogramowania w związku z pojawieniem się nowych standardów oraz narzędzi, które ten proces znacznie upraszczają [2].

2. CHARAKTERYSTYKA USŁUG SIECIOWYCH

Obecnie nie istnieje jednoznaczna i kompletna definicja usług sieciowych. Poniżej przedstawiona definicja podkreśla trzy istotne aspekty usług sieciowych: uniwersalność, powszechność oraz oparcie na standardach – „Usługa sieciowa jest rozproszoną aplikacją sieciową opisaną przy pomocy ustandaryzowanego języka, w obrębie której komunikacja odbywa się przy pomocy standardowych protokołów sieciowych” [3].

Usługi sieciowe można rozpatrywać w kontekście wykorzystywanych obecnie standardów oraz potencjalnych korzyści, co zostało przedstawione w kolejnych podrozdziałach artykułu.

2.1. Standardy

Niemal wszystkie firmy zainteresowane wdrożeniem usług sieciowych dostrzegają potrzebę stworzenia jednolitych standardów. Poniżej zostały przedstawione podstawowe standardy używane obecnie [4]:

- WSDL (ang. *Web Services Description Language*) – język opisu usług sieciowych.
- SOAP (ang. *Simple Object Access Protocol*) – protokół komunikacji.
- UDDI (ang. *Universal Description, Discovery and Integration*) – katalogowanie usług sieciowych.

Wiele podmiotów sygnalizuje potrzebę stworzenia nowych standardów. Istnieje w związku z tym zagrożenie, iż zostanie stworzony cały szereg alternatywnych i niekompatybilnych względem siebie rozwiązań, co może doprowadzić do sytuacji wyjściowej. W chwili obecnej, nie jest możliwa jednoznaczna odpowiedź na pytanie, w jakim kierunku podąży rozwój standardów dla usług sieciowych.

WSDL jest językiem opartym na XML, który opisuje usługi sieciowe oraz sposób dostępu do nich. Podobnie jak XML, WSDL jest językiem „rozszerzalnym” i umożliwia formalny opis usług sieciowych, sposobu implementacji usług sieciowych oraz protokołów komunikacji [5].

Plik WSDL opisuje typy danych niezbędnych do wymiany komunikatów, format komunikatu oraz rodzaj wykonywanej operacji. Pliki WSDL umożliwiają zainteresowanym stronom „podpięcie się” do usługi sieciowej oferowanej przez dany podmiot. Proces ten może być zautomatyzowany dzięki wykorzystaniu narzędzi, które interpretują dokumenty WSDL a następnie tworzą kod w wybranym języku (np. Java), który ułatwia zintegrowanie własnych komponentów systemowych z daną usługą.

SOAP jest prostym protokołem tekstowym opartym na języku XML, który umożliwia wymianę informacji w rozproszonym środowisku. Protokół SOAP został opracowany przez IBM, Microsoft oraz Develop Mentor. SOAP definiuje mechanizm przekazywania komu-

ników oraz parametrów pomiędzy klientami oraz serwerami [6]. Podobnie jak cała koncepcja usług sieciowych, protokół SOAP jest niezależny od m.in. wykorzystywanej platformy aplikacyjnej oraz języka programowania.

UDDI stanowi specyfikację umożliwiającą opisanie usługi sieciowej, udostępnienie tego opisu w sieci oraz odnalezienie usługi sieciowej przez zainteresowanych użytkowników. Implementacja UDDI może mieć postać publicznego repozytorium, w którym przechowywane są informacje nt. dostępnych usług, podmiotów udostępniających dane usługi oraz odsyłacze do nich (adresy plików WSDL) [7]. Publiczne repozytoria UDDI mają charakter rozproszony i są prowadzone przez m.in. IBM, SAP i Microsoft.

Korzystanie z repozytoriów UDDI przypomina przeglądanie książki telefonicznej. Użytkownik poszukujący danej usługi (np. usługi informującej o warunkach pogodowych w danym mieście) może przeszukiwać repozytorium na podstawie wybranych kryteriów (np. słów kluczowych, nazw podmiotów udostępniających usługi) a następnie na podstawie pliku WSDL, którego adres znajduje się w katalogu skorzystać z oferowanej usługi.

Znajdujące się na rynku oprogramowanie umożliwia tworzenie repozytoriów usług o ograniczonym dostępie na prywatny użytek (np. repozytorium dostępne tylko w obrębie sieci lokalnej lub repozytorium dostępne tylko dla stałych partnerów biznesowych).

2.2. Potencjalne korzyści implementacji

Implementacja usług sieciowych wiąże się z osiągnięciem szeregu korzyści. Najważniejsze z nich zostały przedstawione w poniższym zestawieniu.

- Zmniejszenie kosztów integracji oprogramowania.
- Uniwersalność – usługi sieciowe umożliwiają wykorzystanie rozmaitych klientów opartych na różnych językach programowania (np. Java, NET, C++, JavaScript, Perl, itd.). Dla programistów oznacza to, iż cykl życia stworzonych usług będzie zdecydowanie dłuższy od rozwiązań stosowanych do tej pory.
- Łatwość użycia – wykorzystanie usług sieciowych umożliwia udostępnienie w sieci usług, z których interdyscyplinarne zespoły projektowe mogą tworzyć rozwiązanie problemu po stronie klienta poprzez dobór niezbędnych usług. Wykorzystanie istniejących standardów dla usług sieciowych uniezależnia ten proces od języka programowania oraz architektury rozwiązania. Do chwili obecnej zostało stworzonych wiele aplikacji, które umożliwiają niemal automatyczne integrowanie oprogramowania poprzez wykorzystanie usług sieciowych.
- Możliwość wielokrotnego użycia – usługi udostępnione w sieci mogą być używane wielokrotnie, przez różnych partnerów biznesowych oraz łączone w kompleksowe rozwiązania.
- „Czytelne” zarówno dla ludzi jak i komputerów – np. poprzez użycie aplikacji biurowych lub programistycznego interfejsu aplikacji (ang. API – *Application Programming Interface*).
- Powszechnie dostępne – ponieważ usługi sieciowe są dostarczane przez Internet, są dzięki jego infrastrukturze sieciowej dostępne niezależnie od miejsca oraz czasu.
- Stworzone standardy uwzględniają aspekt bezpieczeństwa oraz obecne rozwiązania w tym zakresie (np. firewall).

3. NARZĘDZIA WSPOMAGAJĄCE IMPLEMENTACJĘ USŁUG SIECIOWYCH

Obecnie na rynku istnieje szereg produktów wspierających tworzenie usług sieciowych oraz integrację oprogramowania z ich wykorzystaniem. Narzędzia można podzielić na następujące kategorie:

- Narzędzia wspomagające tworzenie usług sieciowych.
- Narzędzia testujące.
- Narzędzia wspomagające zarządzanie usługami.
- Narzędzia wspomagające korzystanie z usług sieciowych.

3.1. Narzędzia wspomagające tworzenie usług sieciowych

Integracja systemów informatycznych określonej organizacji z systemami jej partnerów biznesowych, w oparciu o standardy usług sieciowych, wymaga odpowiedniej (unifikacji) istniejących wcześniej rozwiązań software'owych. Kluczowym pytaniem w tym kontekście jest pytanie o to jak wykorzystać kod, który został utworzony bez uwzględnienia koncepcji usług sieciowych.

Narzędzia z tej grupy umożliwiają wykorzystanie istniejących komponentów software'owych jako usług sieciowych. Celem ich stosowania jest przyspieszenie procesu integracji oraz minimalizacja kosztów z tym związanych. Na podstawie istniejących komponentów generowany jest kod, dzięki któremu mogą być one wykorzystywane jako usługi sieciowe. Narzędzia z tej grupy różnią się od zintegrowanych środowisk programistycznych (sprzedawanych przez takich producentów jak np. IBM, Borland, Microsoft), gdyż ich celem jest „konwersja” istniejących komponentów na usługi sieciowe a nie tworzenie od podstaw oprogramowania. Do narzędzi z tej grupy można zaliczyć m.in.: Cape Clear Software, Systinet WASP, TierBroker Server oraz Mind Electric GLUE.

3.2. Narzędzia testujące

Testowanie rozproszonych aplikacji, które wykorzystują usługi sieciowe jest znacznie trudniejsze niż w przypadku tradycyjnych rozwiązań scentralizowanych. Sytuację komplikuje dodatkowo fakt, iż usługi sieciowe, z których korzysta dany system mogą być poza kontrolą danego podmiotu.

Kluczowym aspektem działania narzędzi testujących jest możliwość testowania łącznie wielu usług sieciowych (w praktyce najczęściej spotyka się rozwiązania wykorzystujące wiele usług sieciowych naraz). Dla narzędzi z tej grupy istotna jest umiejętność analizy zawartości pakietów SOAP oraz informacji zawartych w pliku WSDL.

Wśród produktów testujących prawidłowość działania usług sieciowych warto wymienić m.in.: Mindreef SOAPscope, Empirix e-TEST Suite, Swingtide oraz TeaLeaf IntegriTea.

3.3. Narzędzia wspomagające zarządzanie usługami

Efektywne zarządzanie usługami sieciowymi stało się istotne ze względu na rozproszone zależności pomiędzy nimi oraz kwestie bezpieczeństwa z tym związane. Narzędzia z tej grupy umożliwiają egzekwowanie ustalonej polityki bezpieczeństwa oraz

sprawne zarządzanie zarówno oferowanymi usługami jak i tymi, z których dany podmiot korzysta (m.in. zarządzanie cyklem życia usług, routing usług, monitorowanie natężenia ruchu, raportowanie, ostrzeganie przed potencjalnymi zagrożeniami).

Do narzędzi wspomagających zarządzanie usługami można zaliczyć m.in.: Actional Looking Glass Management Server and Console, Blue Titan Network Director, Confluent CORE, Infravio Ensemble, Talking Blocks, Digital Evolution oraz Amber Point Management Foundation.

3.4. Narzędzia wspomagające korzystanie z usług sieciowych

Narzędzia z tej kategorii ułatwiają korzystanie z usług udostępnionych przez inne podmioty. Ich rola sprowadza się między innymi do integracji oprogramowania, tworzenia rozszerzeń dla serwisów WWW oraz dostępu dla aplikacji biurowych do użytecznych usług.

Wśród produktów wspomagających korzystanie z usług sieciowych warto wymienić m.in.: Kenamea Composite, Collaxa, Bowstreet Portret Faktory, Grand Central Communications, RatchetSoft oraz M7 Application. Zróżnicowanie funkcjonalności przedstawionych narzędzi jest bardzo duże. Przykładowo M7 generuje programy J2EE, korzystające z wybranych usług, natomiast Bowstreet umożliwia dołączenie usług sieciowych do portali stworzonych przy pomocy IBM WebSphere portals.

4. PRZYKŁADY ZASTOSOWAŃ

W tej części artykułu zostały zaprezentowane przykłady zastosowania usług sieciowych przez następujące firmy:

- Amazon.com [8].
- Life Time Fitness [9].

4.1. Amazon.com

Księgarnia internetowa Amazon.com udostępniła dla wszystkich zainteresowanych pakiet usług sieciowych, które stają się dostępne po procesie rejestracji. Usługi oferowane przez firmę Amazon umożliwiają jej partnerom biznesowym interakcję z serwisem internetowym księgarni przy wykorzystaniu standardowych protokołów sieciowych.

Korzyści, wynikające z wykorzystania usług sieciowych są w tym przypadku następujące:

- Efektywny udział w programie członkowskim (ang Affiliate program). Dowolne podmioty mogą umieścić na swoich stronach przekierowanie do serwisu księgarni. Za każdego zdobytego klienta, który dokona zakupu podmioty otrzymują zapłatę wg przyjętej struktury prowizji. Wykorzystanie usług sieciowych w tym przypadku stanowi wartość dodaną w stosunku do wcześniejszych rozwiązań tradycyjnych i pozwala na większą kontrolę procesów oraz zwiększenie obrotów.
- Na stronach Amazon.com tysiące różnych podmiotów sprzedaje swoje produkty. Poprzez wykorzystanie usług sieciowych sprzedawcy mogą łatwiej kontrolować i zarządzać oferowanym asortymentem, porównywać ceny, itp.

Usługi oferowane przez Amazon.com umożliwiają m.in.: przeszukiwanie bazy danych oferowanych produktów, dodawanie towarów do koszyka zakupów lub sprawdzanie statusu zamówienia.

Aby ułatwić implementację usług partnerom biznesowym Amazon.com udostępnił narzędzie dla programistów które zawiera m.in. podręcznik oraz przykładowe kody dla języków Java, Perl, PHP, SOAP, XML oraz XSLT.

4.2. Life Time Fitness

Life Time Fitness oferuje swoim członkom (ponad 300 tys. zarejestrowanych klientów) sposoby spędzania wolnego czasu w centrach rozlokowanych na terenie Stanów Zjednoczonych (30 ośrodków). Jednym z kluczowych problemów, z którym klub się zmagał była efektywna obsługa członków (m.in. zapisy na indywidualne zajęcia, zapisy na masaż, rezerwacja kortów tenisowych oraz innych obiektów sportowych). Do roku 2002 rezerwacje były realizowane telefonicznie, co było niezwykle czasochłonne.

W 2002 roku klub zaoferował internetowy portal zbudowany przy wykorzystaniu usług sieciowych. Obecnie użytkownicy po zalogowaniu mogą efektywnie zaplanować dowolne zajęcia. Pracownicy klubu mogą natomiast w wygodny sposób zarządzać harmonogramami oraz monitorować rezerwacje klientów. Usługi sieciowe umożliwiają efektywną wymianę danych pomiędzy interfejsem użytkownika umieszczonym na stronie WWW a wewnętrzną rozproszoną aplikacją zarządzającą harmonogramami. Wdrożone rozwiązanie przyniosło firmie wymierne korzyści finansowe. Dostrzegając te zalety około 80 różnych organizacji (m.in. klubów fitness, uczelni) zadeklarowało chęć zakupu powyższego systemu oraz jego adaptacji do własnych potrzeb.

5. ZAKOŃCZENIE

Aby implementacja usług sieciowych przyniosła wymierne korzyści, musi zostać poprzedzona szczegółową analizą potrzeb. Należy pamiętać, iż implementacja sama w sobie wiąże się z poniesieniem kosztów, których zdyskontowanie jest zazwyczaj możliwe dopiero w dłuższym okresie czasu.

Decydując się na wykorzystanie usług sieciowych należy pamiętać także o pewnych niedogodnościach związanych z ich wykorzystaniem (np. brak trwałych standardów, brak kompletnego rozwiązania w obszarze bezpieczeństwa). Pomimo pewnych minusów usługi sieciowe są postrzegane jako przyszłość w obszarze integracji oprogramowania, zautomatyzowanej wymianie informacji z partnerami biznesowymi, w łatwej rozbudowie systemów oraz szeroko pojętych rozwiązaniach e-biznesowych.

Usługi sieciowe z pewnością będą odgrywały coraz większą rolę w najbliższych latach, dlatego już teraz warto zastanowić się nad zasadnością ich wykorzystania oraz możliwościami ich włączenia w firmową strategię informatyzacji.

BIBLIOGRAFIA

- [1] Pyrovolakis O., Garbi A., Plataniotis A.: *Web-Service Framework for Business Process Modelling & Legacy Systems Integration*. W: Proceedings of the eChallenges Conference. October 2002, Bologna, Italy.
- [2] *IBM Web Services for On Demand e-business. Maximizing Opportunities Today Through the Industry's Broadest Support for Web Services*, IBM Software Group, 2003.
- [3] Curbera F., Nagy W., Weerawarana S.: *Web Services: Why and How*, IBM T.J. Watson Research Center, 2001.
- [4] Kreger H.: *Web Services Conceptual Architecture (WSCA 1.0)*, IBM Software Group, 2001.
- [5] <http://www.w3.org/TR/wsdl>.
- [6] <http://www.w3.org/TR/SOAP/>.
- [7] <http://www.uddi.org>.
- [8] <http://Amazon.com>.
- [9] Patton S.: *Web Services in the Real World*, CIO Magazine, April 2002.

INTRODUCTION TO WEB SERVICES – STANDARDS, IMPLEMENTATION BENEFITS AND CASE STUDIES

Summary

Within the last years, it becomes crucial for company business activities to integrate its software inside the company as well as with software used by business partners. Methods developed by independent companies and used to integrate various software components will lose in the long term its significance, because they require high programming skills and are usually time-consuming. The solution of this problem might be the use of so-called Web Services in order to integrate software. The aim of this paper is to present the essence of Web Services as regards standards, areas of application and benefits. The paper discusses also tools supporting implementation of Web Services and case studies.

**Paweł Czarnul, Michał Bajor, Anna Banaszczyk, Paweł Buszkiewicz,
Marcin Fiszer, Marcin Frączak, Michał Klawikowski, Jacek Rakiej,
Katarzyna Ramczykowska, Krzysztof Suchcicki**

**Katedra Architektury Systemów Komputerowych
Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
WWW: <http://www.ask.eti.pg.gda.pl>, <http://fox.eti.pg.gda.pl/~pczarnul>
e-mail: pczarnul@eti.pg.gda.pl, tel. (058) 347 25 24, fax. (058) 347 28 63**

ROZPROSZONE USŁUGI OBLICZENIOWE NA KLASTRACH TASK Z DOSTĘPEM PRZEZ WWW I „WEB SERVICES”^{1,2}

Streszczenie

Artykuł omawia architekturę i doświadczenia wstępnych faz specyfikacji wymagań i analizy projektu, który umożliwi zdalne wykorzystywanie równoległych klastrów i/lub superkomputerów sieci TASK (i innych) przez użytkowników z rozproszonych geograficznie lokalizacji. System zapewni dotychczasowym i nowym użytkownikom klastrów i superkomputerów TASK zdalne uruchamianie i zarządzanie aplikacjami, bibliotekami równoległymi i sekwencyjnymi poprzez przyjazny interfejs WWW i/lub Web Services (usługi obliczeniowe). Architektura systemu przewiduje niezależne serwery J2EE obsługujące warstwę prezentacyjną i logiczną systemu oraz odpowiednią warstwę komunikacyjną do zarządzania klastrami, zarówno z jak i bez systemów kolejkowych. System nie wymaga modyfikacji oprogramowania ani struktury kont użytkowników na poszczególnych klastrach.

1. WSTĘP

Rozwój systemów równoległych i rozproszonych wykorzystujących niskopoziomowe biblioteki i narzędzia jak MPI i PVM ([1], [2], [3]) do symulacji skomplikowanych zjawisk fizycznych ([4-6]) pociąga za sobą konieczność stworzenia intuicyjnych interfejsów/systemów zarządzania aplikacjami równoległymi i pracą grupową w środowiskach rozproszonych ([7]). Systemy typu grid ([8-13]) umożliwiają kontrolowane współdzielenie zasobów pomiędzy ośrodkami na międzynarodową skalę. Jednakże ich rozwój, ze względu na wielkość i stopień skomplikowania, cechuje się trudnością szybkiego dostosowania, w szczególności do najnowszych technologii dostępu do zasobów. Dlatego też w niniejszej

¹ obliczenia wykonano na komputerach Centrum Informatycznego Trójmiejskiej Akademickiej Sieci Komputerowej

² częściowo wykonano w ramach grantu KBN 4 T11C 00525

pracy proponujemy architekturę (będącego aktualnie w fazie projektu) systemu mniejszej skali, który umożliwi zdalne wykorzystanie równoległych klastrów oraz superkomputerów sieci TASK przez użytkowników z rozproszonych geograficznie lokalizacji. System zapewni dotychczasowym i nowym użytkownikom klastrów i superkomputerów TASK zdalne uruchamianie i zarządzanie aplikacjami, bibliotekami równoległymi i sekwencyjnymi poprzez przyjazny interfejs WWW i/lub Web Services ([14-15]). Dostęp do systemu odbywał się będzie przez serwery J2EE ([16], rysunek 1) będące pojedynczymi punktami dostępu do klastrów. Architektura uniezależnia serwery dostępu od klastrów i mechanizmów/oprogramowania wykorzystywanych wewnątrz nich dzięki komunikacji poprzez zdalne (ukryte przed użytkownikiem) wywołania ssh do klastrów. Pozwala to na zaprojektowanie i implementację szeregu wysokopoziomowych funkcji niezależnie od systemów kolejkowych, wykorzystywanej powłoki itd. Jednocześnie każdy użytkownik, który ma już konto na klastrach (co wymaga odpowiednich procedur administracyjnych nie mieszczących się w gestii twórców czy instalatorów systemu omawianego tutaj), może wykorzystywać w łatwy sposób (poprzez WWW oraz dodatkowo za pomocą serwera Web Services) dostępne usługi po jednorazowym zarejestrowaniu loginu/hasła do klastrów, które chce używać poprzez opisywany system a następnie używanie pojedynczego loginu/hasła do systemu. Co więcej, niezależność od klastrów umożliwia zaprojektowanie zwielokrotniania usług dostępu poprzez instalację niezależnych serwerów J2EE. Poprawia to czas reakcji dla użytkownika (przy dużym obciążeniu) jak i daje niezawodność (możliwość korzystania z innego serwera po awarii, gdyż stan sesji może być przechowywany w bazie danych – podobne rozwiązania stosowane są np. na platformie .NET – [17]).

2. SPECYFIKACJA FUNKCJONALNA

W ramach systemu udostępniane są następujące usługi:

1. Możliwość korzystania z zainstalowanych na klastrze aplikacji/bibliotek udostępnionych przez danego lub innych użytkowników – łatwe wywołanie poprzez przeglądarkę WWW lub Web Services ([14-15]) z uwzględnieniem logicznego przyporządkowania procesów aplikacji równoległych do węzłów klastrów. Definiuje się tutaj grupy robocze, w ramach których udostępnione są wywołania konkretnych aplikacji lub bibliotek. Zakłada się dynamiczne dodawanie oraz usuwanie aplikacji lub bibliotek.

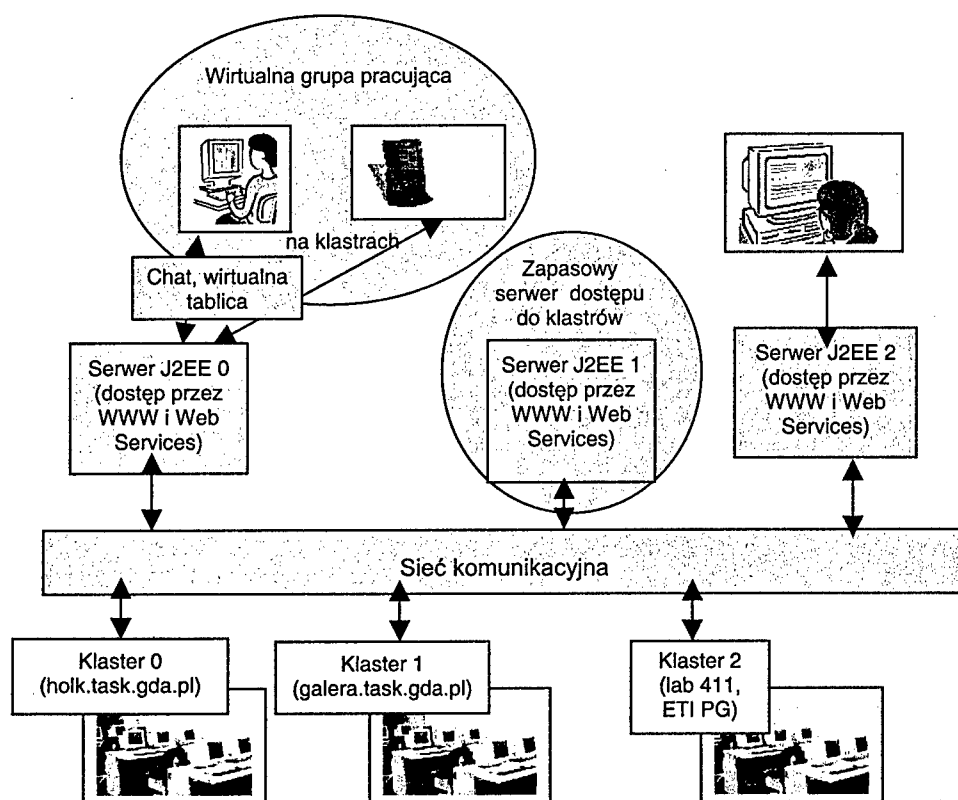
Ponadto przewidziana została możliwość (również odpłatnego – system licencjonowania) uruchamiania aplikacji udostępniona innym użytkownikom.

Interfejs pozwoli na: intuicyjne zarządzanie danymi usług obliczeniowych w tym łatwe ponowne wywołania dla nowych danych wejściowych przez WWW/Web Services oraz łatwe wywoływanie (zarówno z wykorzystaniem systemu kolejkowego jak i interaktywnego wykonania aplikacji równoległej/sekwencyjnej) oraz śledzenie postępu wykonania, zarówno na klastrach wykorzystujących systemy kolejkowe jak i ogólnie dostępnych sieci komputerowych dostępnych do pracy interaktywnej. Pośrednie oraz finalne wyniki będą mogły być oglądane w przeglądarce. System umożliwi tworzenie skryptów zawierających sekwencje zdalnego kopiowania/kompilacji/wywołań/wyświetlenia wyników aplikacji wykonywanych na klastrach.

W ramach testów planowane jest uruchomienie (oraz przetestowanie i zainstalowanie jako dostępnych usług obliczeniowych w systemie) przykładowych aplikacji równoległych w paradygmatach SPMD, dziel-i-zwyciężaj itd.

Użytkownicy będą bezpiecznie uwierzytelniani do własnej sesji z poziomu przeglądarki WWW i/lub klientów zaimplementowanych w wielu językach programowania wykorzystujących Web Services – baza danych użytkowników.

2. Dostęp do własnych zasobów plikowych na klastrze – możliwość kopiowania do przestrzeni dyskowej na klastrze z własnego komputera jak również plików pomiędzy klastrami z możliwym wsparciem dla aplikacji w celu wykorzystania przestrzeni wielu klastrów do przechowywania danych i ich automatycznego równoważenia.
3. Wizualizacja obciążenia na poszczególnych węzłach klastra poprzez WWW.
4. Komunikacja w czasie rzeczywistym pomiędzy użytkownikami/członkami zdefiniowanych grup za pomocą przeglądarek WWW.



Rys. 1. Schemat przetwarzania proponowanego systemu

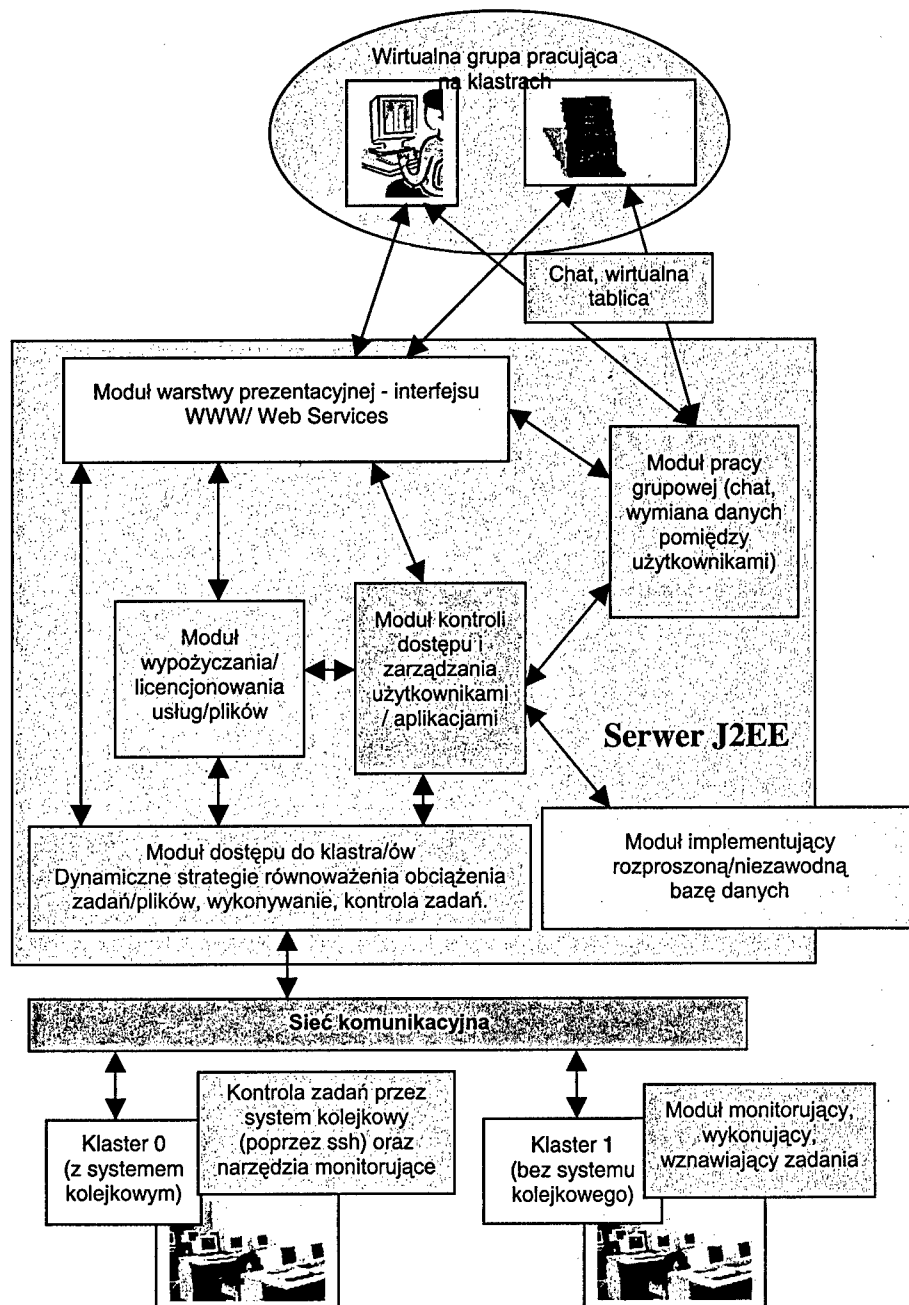
3. ARCHITEKTURA SYSTEMU

Architektura proponowanego systemu jest elastyczna i pozwala na scalenie wielu rozproszonych klastrów z dostępem przez ww. serwery J2EE w jeden spójny system, na wzór systemów typu grid ([8-13]), zorientowanym jednak ściśle na publikację aplikacji/bibliotek klastrów przez WWW/WebServices. Architektura systemu zaprojektowana została

w ten sposób, aby możliwe było zintegrowanie równoległych, rozproszonych geograficznie klastrów obliczeniowych wraz z mechanizmami pracy zespołowej. System będzie wówczas umożliwiał kolejikowanie zgłoszeń uruchomienia aplikacji równoległych DAMPVM, ([18-20]), PVM ([1]) i MPI ([2]) i innych od rozproszonych geograficznie użytkowników. Dana aplikacja uruchamiana jest na najlepszym (z punktu widzenia podanego kryterium np. minimalnego czasu wykonania lub maksymalnej dostępnej przestrzeni dyskowej) klastrze. W zależności od dostępnego oprogramowania jak np. DAMPVM, PVM, MPI możliwe jest wówczas przetwarzanie równoległe oraz wielowątkowe w ramach klastrów i węzłów. Architektura dopuszcza implementację rozproszonego systemu plików, który potrafiłby partycjonować pliki dużych rozmiarów a następnie przechowywać w sposób rozproszony i bezpieczny, również z replikacją procesów i równoważeniem obciążenia ([21-23]. Wymagające aplikacje równoległe ([5-6]) uruchamiane na klastrach wymagają bowiem z reguły ogromnych ilości danych, tymczasowych lub końcowych, które wykraczają poza pojemność dyskową nawet całego klastra. Tyczy się to w szczególności symulacji zjawisk fizycznych ([5-6]) itp.

W fazie analizy/projektu systemu ustalono wykorzystanie następujących narzędzi/technologii do implementacji poszczególnych warstw systemu:

1. Warstwa prezentacyjna i logiczna: serwery J2EE (równoważne funkcjonalnie) umożliwiające użytkownikowi dostęp przez WWW/usługi sieciowe (Web Services) do zasobów na klastrach poprzez wybieranie najlepszego klastra do wykonania zadania, łatwe w obsłudze przesyłanie plików pomiędzy kontami różnych klastrów. Serwery J2EE są niezależne pod względem przetwarzania tj. umożliwiają zwiększenie wydajności systemu poprzez równoległą obsługę wielu zgłoszeń od klientów jak również zapewniają niezawodność – awaria jednego serwera dostępowego umożliwia dalszą pracę poprzez inny, który korzystał będzie z tej samej, rozproszonej i niezawodnej bazy danych.
 - a. Warstwa prezentacyjna systemu to serwlety oraz strony JSP udostępnione z poziomu serwera J2EE jak również równoważne funkcjonalnie usługi sieciowe (Web Services), w systemie warstwa ta implementowana jest przez (rysunek 2):
 - i. moduł warstwy prezentacyjnej – interfejsu WWW/Web Services,
 - ii. moduł pracy grupowej (chat, wymiana danych pomiędzy użytkownikami)
 - b. Warstwa logiczna systemu składa się z kilku modułów:
 - i. moduł kontroli dostępu i zarządzania użytkownikami/aplikacjami,
 - ii. moduł wypożyczania/licencjonowania usług/plików,
 - iii. moduł dostępu do klastrów – dynamiczne strategie równoważenia obciążenia zadań/plików, wykonywanie, kontrola zadań – moduł ten komunikuje się za pomocą zdalnych wywołań ssh z poszczególnymi klastrami, co uniezależnia system od konkretnych klastrów,
 - iv. moduł implementujący rozproszoną i niezawodną bazę danych – wychodzi poza granice serwera J2EE, implementując bazę danych na wielu dostępnych komputerach, udostępniając standardowy sterownik JDBC innym modułom systemu. Moduł ten wykorzystywany jest przez wiele serwerów J2EE umożliwiających dostęp do klastrów.



Rys. 2. Architektura systemu: moduły serwera J2EE i komunikacja pomiędzy nimi

2. Warstwa klastrów. Wyróżnia się tutaj:

- a. Klastry z dostępem przez system kolejkowy np. LSF, PBS itp., dla których opracowany zostanie protokół wymiany komunikatów poprzez zdalne wywołania ssh do uruchamiania zadań (również odpłatne uruchamiania przez innych użytkowników, jeśli zostało dozwolone), kopiowania plików itp.,
- b. Klastry wolnodostępnych komputerów np. laboratoriów, gdzie komputery mogą być bez ostrzeżenia wyłączane. W ramach systemu zaimplementowany zostanie moduł, który kontrolował będzie stan takiego klastra, w razie potrzeby uruchamiając ponownie procesy kontroli klastra i wznowiając wykonanie zadań.

W ramach klastrów przewiduje się instalację i uruchamianie aplikacji/bibliotek typowych dla klastrów: specjalizowanych aplikacji użytkowników – zastosowania w chemii, fizyce itp., DAMPVM ([18-20]), PVM ([1]), MPI ([2]) do wydajnej komunikacji wewnątrz klastrów.

4. DODATKOWE WYMAGANIA I OPROGRAMOWANIE

System opracowywany jest z myślą o umożliwieniu łatwego dostępu użytkownikom do klastrów TASK (np. galera.task.gda.pl, holk.task.gda.pl), którzy posiadają już konta. Wymagało to będzie jednokrotnej rejestracji w dowolnym serwerze J2EE systemu oraz dowolnej przeglądarki WWW, aby móc korzystać z łatwego uruchamiania zadań oraz zarządzania plikami poprzez WWW i usługi sieciowe (Web Services), nie tylko na jednym, ale na wielu połączonych klastrach. Ponadto system umożliwił będzie zdalną pracę grupową a nawet udostępnianie aplikacji (jeśli zgodne z regulaminem klastrów), nawet odpłatne, innym.

Rozwijając architekturę systemu PVMWebCluster ([7-9]) i innych ([24]), gdzie wykorzystane były serwery WWW, Tomcat, AXIS ([14-15]), MICO CORBA ([25]) oraz PVM, MPI ([1], [2]), tutaj cała strona prezentacyjna i logiczna systemu (serwery J2EE oraz wszystkie moduły wykraczające poza jego granice) zaimplementowane zostaną za pomocą darmowych narzędzi i technologii, głównie w języku Java.

Na poziomie klastrów, celowe wydaje się wykorzystanie dostępnych narzędzi do śledzenia wykonania aplikacji (w szczególności równoległych) jak i programów umożliwiających graficzną ocenę wydajności wykonania aplikacji, co następnie mogłoby być przedstawione użytkownikowi poprzez interfejs WWW. Tutaj rozważyć należy użycie szeregu narzędzi takich jak lub podobnych do: Totalview, AIMS, Dyninst, Pablo, nupshot, Paradyn, WAMPIR, PAPI, PARDON ([26-27]). Ponadto system mógłby umożliwiać łatwy dostęp do graficznych interfejsów aplikacji pozwalających na wykonanie symulacji poprzez dostępne pakiety np. programu Matlab jak toolboxy do sieci neuronowych, algorytmów genetycznych, symulacji układów elektronicznych czy innych. Dostępność takich narzędzi pomogłaby przeprowadzić analizę wykonania aplikacji równoległej, co jest często niezbędne w symulacjach na klastrach oraz zwiększyć funkcjonalność systemu.

5. ZAKOŃCZENIE

Efektem projektu będzie fizyczna implementacja opisywanego systemu. Dotychczasową pracę badawczą zespołu KASK ETI PG proponującego projekt (publikacje na konferencjach EuroPVMMPI, PDCS, PARELEC, PPAM i innych) stanowią pokrewne prototypowe rozwiązania. Planujemy udostępnienie systemu dotychczasowym i nowym

użytkownikom TASK. W ramach testów wykonane zostaną testy wydajnościowe na klastrach przykładowej biblioteki równoległych funkcji operacji na obrazach graficznych udostępnionej następnie przez WWW/Web Services wraz z różnymi parametrami wykorzystania klastrów jak liczba procesorów na węzeł, logiczne przyporządkowanie zadań do procesorów itp. Planowane jest utworzenie dedykowanego serwisu WWW poświęconego systemowi, zawierającego wersję instalacyjną systemu, dokumentację i podręcznik użytkownika. Planuje się również wykorzystanie zaimplementowanego systemu w Studium Podyplomowym „Zaawansowane aplikacje i usługi internetowe” uruchomionym w roku akademickim 2003/2004 na wydziale ETI Politechniki Gdańskiej oraz przez studentów Wydziału ETI Politechniki Gdańskiej do zajęć z przetwarzania równoległego i rozproszonego.

BIBLIOGRAFIA

- [1] Geist, A., Beguelin, A., Dongarra, J., Jiang, W., Manchek, R., Sunderam, V.: *PVM Parallel Virtual Machine. A Users Guide and Tutorial for Networked Parallel Computing*. MIT Press, Cambridge (1994) <http://www.epm.ornl.gov/pvm/>.
- [2] Wilkinson, B., Allen, M.: *Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers*. Prentice Hall (1999)
- [3] Buyya, R., ed.: *High Performance Cluster Computing, Programming and Applications*. Prentice Hall (1999)
- [4] Czarnul, P., Tomko, K., Krawczyk, H.: *Dynamic Partitioning of the DivideandConquer Scheme with Migration in PVM Environment*. In: Recent Advances in Parallel Virtual Machine and Message Passing Interface. Number 2131 in Lecture Notes in Computer Science, Springer-Verlag (2001) 174--182 8th European PVM/MPI Users' Group Meeting, Santorini/Thera, Greece, September 23-26, 2001, Proceedings.
- [5] Costas D. Sarris, Pawel Czarnul, Senthil Venkatasubramanian, Werner Thiel, Karen Tomko, Linda P. B. Katehi, Barry S. Perlman: "Mixed Electromagnetic - Circuit Modeling and Parallelization for Rigorous Characterization of Cosite Interference in Wireless Communication Channels" at DoD/HPCMO Users Conference, U.S.A., 2002, <http://www.hpcmo.hpc.mil/Htdocs/UGC/UGC02/index.html>
- [6] Costas D. Sarris, Karen Tomko, Pawel Czarnul, Shih-Hao Hung, Robert L. Robertson, Donghoon Chun, Edward S. Davidson, Linda P. B. Katehi: "Multiresolution Time Domain Modeling for Large Scale Wireless Communication Problems" in Proceedings of the 2001 IEEE AP-S International Symposium on Antennas and Propagation vol. 3, pages 557-560, 2001 Foster, I., Kesselman, C., eds.: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann (1998) ISBN 1558604758.
- [7] Czarnul, P.: *PVMWebCluster: Intergration of PVM Clusters Using Web Services and CORBA*. In: Recent Advances in Parallel Virtual Machine and Message Passing Interface. Number 2840 in Lecture Notes in Computer Science (2003) 268-275
- [8] Foster, I., Kesselman, C., Tuecke, S.: *The Anatomy of the Grid: Enabling Scalable Virtual Organizations*. International Journal of High Performance Computing Applications 15 (2001) 200--222 <http://www.globus.org/research/papers/anatomy.pdf>.
- [9] Foster, I., Kesselman, C., Nick, J., Tuecke, S.: *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. In: Open Grid Service Infrastructure WG. (2002) Global Grid Forum, <http://www.globus.org/research/papers/ogsa.pdf>.
- [10] *Globus System* (<http://www.globus.org>)
- [11] *EuroGrid* (<http://www.eurogrid.org>)
- [12] *GridLab* (<http://www.gridlab.org>)
- [13] Roman Wyrzykowski, koordynator: *CLUSTERIX*, <http://clusterix.pcz.pl>

- [14] Streicher, M.: Creating Web Services with AXIS: *Apache's Latest SOAP Implementation Bootstraps Web Services*. Linux Magazine (2002) http://www.linuxmag.com/200208/axis_01.html.
- [15] Butek, R., Chappell, D., Daniels, G., Davis, D., Haddad, C., Jordahl, T., Loughran, S., Nakamura, Y., Ruby, S., Rineholt, R., Sandholm, T., Scheuerle, R., Sedukhin, I., Seibert, M., Sitze, R., Srinivas, D.: *Axis User's Guide*. 1.1 edn. (2001) <http://cvs.apache.org/viewcvs.cgi/~checkout~/xmlaxis/java/docs/userguide.html>.
- [16] Sun Microsystems: *Java 2 Platform, Enterprise Edition (J2EE)*, <http://java.sun.com/j2ee>
- [17] J. Prosise: *Programming Microsoft .NET*, Microsoft Press, 2002, ISBN 0-7356-1376-1
- [18] Czarnul, P.: *Programming, Tuning and Automatic Parallelization of Irregular Divideand-Conquer Applications in DAMPVM/DAC*. International Journal of High Performance Computing Applications 17 (2003) 77--93
- [19] Czarnul, P., Krawczyk, H.: *Dynamic Assignment with Process Migration in Distributed Environments*. In: *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Number 1697 in Lecture Notes in Computer Science (1999) 509--516
- [20] Czarnul, P.: *Dynamic Process Partitioning and Migration for Irregular Applications*. In: International Conference on Parallel Computing in Electrical Engineering PARELEC'2002, Proceedings, Warsaw, Poland (2002) <http://www.parelec.org>.
- [21] Fagg, G.E., Gabriel, E., Resch, M., Dongarra, J.J.: *Parallel IO Support for Metacomputing Applications: MPI Connect IO Applied to PACXMPI*. In: *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Number 2131 in Lecture Notes in Computer Science, Springer-Verlag (2001) 135--147 8th European PVM/MPI Users' Group Meeting, Santorini/Thera, Greece, September 23-26, 2001, Proceedings.
- [22] Alexandrov, A.D., Ibel, M., Schauser, K.E., Scheiman, C.J.: *Extending the Operating System at the User Level: the Ufo Global File System*. In: Proceedings of the USENIX Annual Technical Conference, Anaheim, California, USA (1997) 77--90
- [23] Rhea, S., Wells, C., Eaton, P., Geels, D., Zhao, B., Weatherspoon, H., Kubiawicz, J.: *MaintenanceFree Global Data Storage*. IEEE Internet Computing 5 (2001) 40--49
- [24] Jorba, J., Bustos, R., Casquero, A., Margalef, T., Luque, E.: *Web Remote Services Oriented Architecture for Cluster Management*. In: *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Number 2474 in LNCS, Springer-Verlag (2002) 368--375 9th European PVM/MPI Users' Group Meeting, Linz, Austria, Sept/Oct 2002, Proceedings.
- [25] MICO CORBA, <http://www.mico.org>
- [26] *Debugging and Tuning Tools, Performance Tools*, <http://www.lanl.gov/projects/asci/bluemtn/software/debugging.htm>
- [27] University of Wisconsin, USA: *Paradyn Parallel Performance Tools*, <http://www.cs.wisc.edu/~paradyn/>

DISTRIBUTED COMPUTING SERVICES ON TASK CLUSTERS ACCESSIBLE VIA WWW AND WEB SERVICES

Summary

The paper presents the architecture and experiences during the early stages of requirements specification and analysis of a project which enables to make use of distributed (TASK and others) clusters by geographically distant clients. The system will allow existing and new users remote job submission, application and library management via easy-to-use WWW and Web Services interfaces. The architecture consists of separate J2EE servers handling the presentation and the logic layers of the system as well as supporting an adequate communication infrastructure to communicate with clusters with and without queueing systems. The system does not require any modifications in the account/cluster management making its deployment easy.

Madian dit Tiéman Diarra, Agnieszka Gwoździńska, Jerzy Kaczmarek

Katedra Inżynierii Oprogramowania, WETI, Politechnika Gdańska

WYKORZYSTANIE SERWERÓW UDDI DLA SYSTEMÓW ZDALNEJ EDUKACJI

Streszczenie

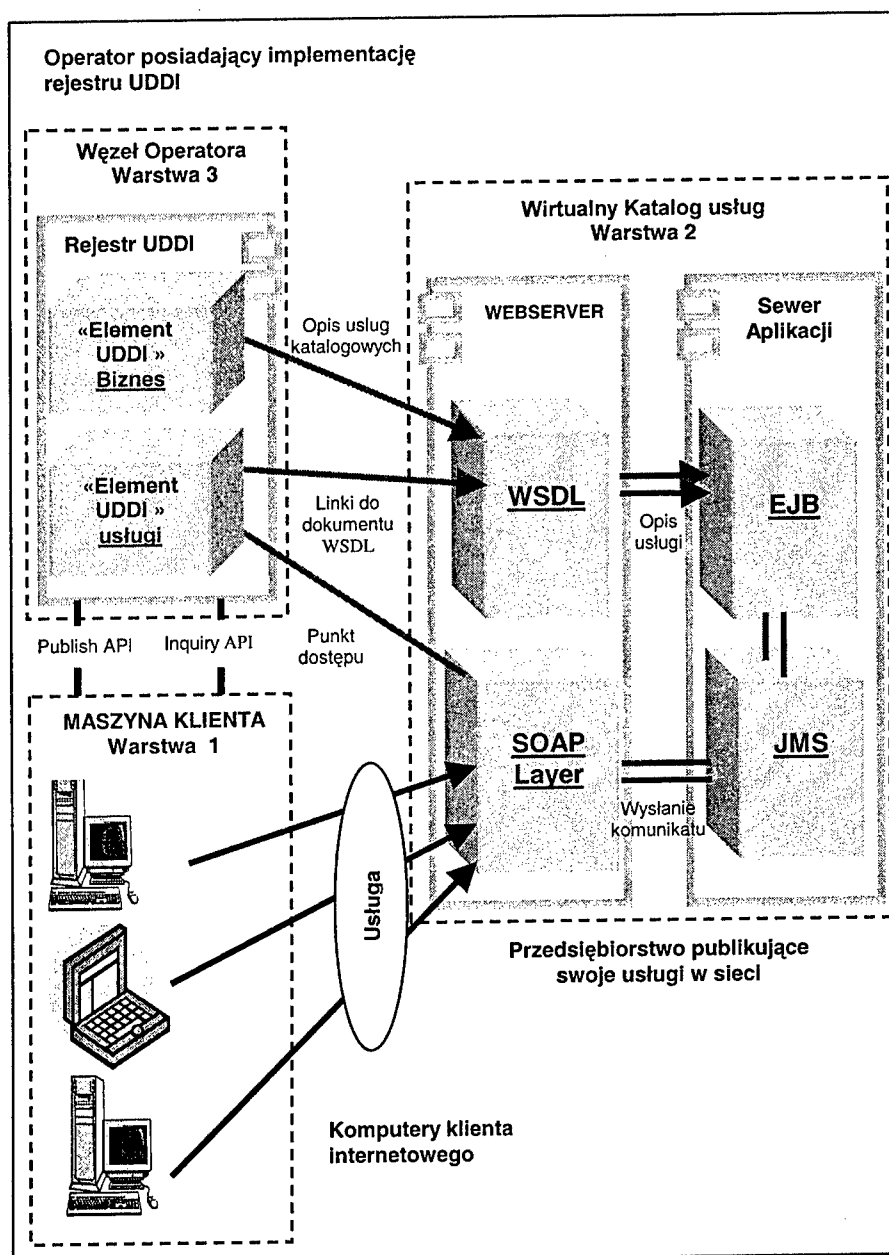
Sieć Internet zawiera obecnie znaczącą ilość materiałów edukacyjnych. Stanowią one zwykle jednolitą całość w postaci kompletnych wykładów. Współczesne tendencje w e-learningu zmierzają w kierunku podzielenia materiałów edukacyjnych na mniejsze części zwane obiektami edukacyjnymi. W artykule przedstawiono możliwości wykorzystania usług sieciowych opartych na serwerach UDDI (*Universal Description, Discovery and Integration*) do przechowywania obiektów edukacyjnych. Opisane rozwiązanie wykorzystujące język XML, protokół SOAP (*Simple Object Access Protocol*) i standard WSDL (*Web Services Description Language*) pozwala na tworzenie baz wiedzy przeznaczonych zarówno dla uczniów jak również dla nauczycieli. Umożliwia nauczycielom poszukiwanie, modyfikowanie i wykorzystywanie istniejących materiałów multimedialnych. W pracy wykazano, że daje to możliwość organizowania procesu dydaktycznego na wyższym poziomie jakości.

1. WSTĘP

Rozwój Internetu doprowadził do powszechnego wykorzystywania przez użytkowników informacji w nim zawartych. Wśród dużej liczby różnych typów danych na uwagę zasługują informacje edukacyjne [1]. Obecnie stanowią one zwykle jednolitą całość w postaci lekcji czy wykładów. Współczesne trendy w e-learningu zmierzają w kierunku podziału materiałów dydaktycznych na obiekty edukacyjne. Istnieją już międzynarodowe standardy definiujące kształt takiego obiektu. Powszechne wykorzystanie rozproszonych w sieci multimedialnych materiałów edukacyjnych wymaga zaprojektowania technologii, która umożliwi stworzenie globalnej sieci połączonych i współpracujących z sobą aplikacji. Zgodnie z tą technologią każda aplikacja traktowana będzie jako usługa sieciowa, która udostępnia swoją funkcjonalność poprzez interfejs programowy. Dlatego niezbędne staje się wykorzystywanie wspólnych standardowych mechanizmów. Do tworzenia obiektów edukacyjnych powinien być wykorzystywany język XML, do przesyłania danych między komputerami protokół SOAP a do prezentacji danych w interfejsie użytkownika standard WSDL. Nowoczesne technologie i usługi Internetowe mogą podnieść na wyższy poziom jakość nauczania w wielu krajach.

2. USŁUGI SIECIOWE OPARTE O SERWERY UDDI

Na rysunku 1 przedstawiono wszystkie elementy trójwarstwowego systemu realizującego usługę sieciową polegającą na poszukiwaniu, przekazywaniu i prezentowaniu danych podzielonych na obiekty o znanej lokalizacji, w tym także obiekty edukacyjne [2].



Rys.1. Architektura usług sieciowych

Warstwa pierwsza znajduje się na komputerze użytkownika, ucznia czy nauczyciela. Jest odpowiedzialna za prezentację danych oraz interakcję z użytkownikiem. Użytkownik zwykle przez przeglądarkę lub dedykowaną aplikację wysyła żądanie do serwera usług i otrzymuje informacje często multimedialne w postaci stron HTML.

Warstwa druga stanowi serwer aplikacji i usług, z którym użytkownik komunikuje się poprzez protokół SOAP [3,4]. Protokół ten jest bardziej zaawansowany niż tradycyjny protokół, HTML. Umożliwia tworzenie złożonych żądań i odpowiedzi, scenariuszy, komunikację konwersacyjną i budowę wielowęzłowych dróg komunikacji. Protokół SOAP jest niezależny od platformy sprzętowej i języka programowania. Komunikat SOAP składa się z trzech części: koperty, nagłówka i treści. Koperta dostarcza opakowania dla komunikatu, wewnątrz koperty znajdują się nagłówek, treść i informacje o błędach.

Usługi sieciowe nie mają interfejsu użytkownika. Programowy interfejs usług zapewnia język WSDL. Wspiera on większość języków programowania, opisuje typy danych, format komunikatu oraz rodzaj wykonywanej operacji.

Warstwę trzecią stanowi serwer UDDI. Serwer UDDI jest rodzajem rejestru, w którym umieszcza się informacje o rodzaju obiektu i miejscu, gdzie się znajduje w rozproszonej sieci komputerowej.

Ideą serwerów UDDI miała być otwarta inicjatywa przemysłowa pozwalająca przedsiębiorstwom na wzajemne poszukiwanie, definiowanie strategii współdziałania i dzielenie się posiadanymi informacjami w globalnej architekturze rejestrowej. Miała stać się centralnym katalogiem podobnym do książki telefonicznej. Istnieją dwa przypadki użycia rejestru UDDI, wyszukiwanie i publikowanie.

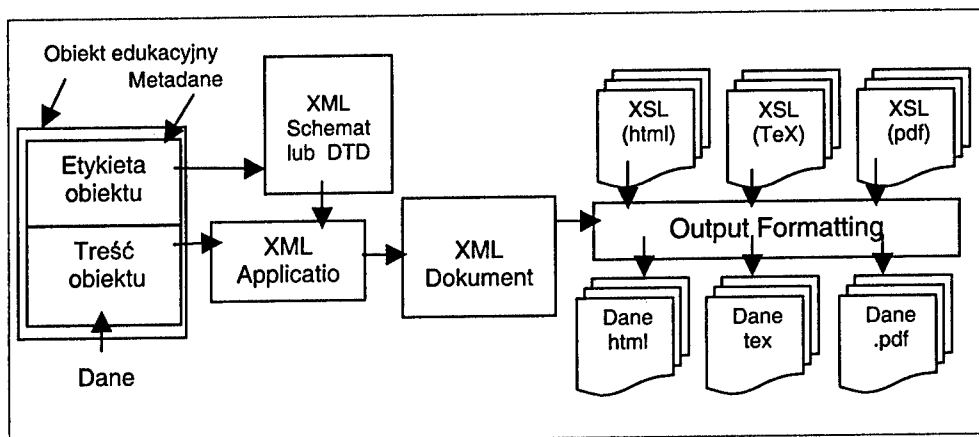
Przy wyszukiwaniu wykorzystuje się mechanizmy alokacji danych na podstawie nazwy słów kluczowych lub typu usługi. Publikowanie wymaga standaryzowanego modelu danych, który zawiera informacje o dostawcy, usłudze i miejscu lokalizacji obiektu w rozproszonej sieci serwerów.

W rejestrach UDDI można umieszczać nie tylko obiekty przemysłowe, ale również inne typy obiektów takie jak obiekty edukacyjne. Stwarza to możliwość wykorzystania tej technologii w stworzeniu wirtualnej przestrzeni edukacyjnej zawierającej rozproszone obiekty edukacyjne.

3. OBIEKT EDUKACYJNY

Obiektem edukacyjnym może być każdy cyfrowy zbiór danych wykorzystywany wielokrotnie w procesie nauczania. Współczesne trendy w dziedzinie zdalnego nauczania zmierzają do podziału materiałów edukacyjnych na mniejsze części zwane obiektami edukacyjnymi. Każdą lekcję, książkę czy inny materiał edukacyjny można podzielić na mniejsze, logiczne części stanowiące spójną całość będącą obiektem edukacyjnym [5]. Takie obiekty mogą być przechowywane z wykorzystaniem serwerów UDDI i stać się przedmiotem poszukiwania i przekazywania w ramach usług Internetowych.

Na rysunku 2 przedstawiono sposób wykorzystania obiektu edukacyjnego w aplikacji klienta.



Rys.2. Wykorzystanie obiektu edukacyjnego

Budowa obiektu edukacyjnego musi być zgodna z wymaganiami technologii usług internetowych. Obiekt edukacyjny zgodnie ze standardami musi składać się oprócz treści z metadanych umieszczonych w tak zwanym manifeście. Manifest zawiera ważne informacje opisujące obiekt takie jak na przykład słowa kluczowe, na podstawie których można poszukiwać obiektów o zadanej treści. Obiekty edukacyjne muszą być tworzone w oparciu o język XML, który oddziela treści danych od sposobu ich prezentacji. Separacja treści od rodzaju formatowania pozwala wykorzystywać obiekt edukacyjny przez dowolne aplikacje klienckie. Jak pokazano na rys.2 dokument XML jest po stronie klienta formatowany i tworzona jest multimedialna strona możliwa do wyświetlenia przez przeglądarkę.

Wykorzystanie obiektów edukacyjnych, które będą miały standardowy format pozwoli na budowanie kursów edukacyjnych w oparciu o obiekty już istniejące, a stworzone przez różne inne podmioty. Daje to możliwość produkcji materiałów edukacyjnych o wysokiej jakości użytkowej i merytorycznej.

4. BAZA WIEDZY DLA NAUCZYCIELI

Nauczyciele często tworzą materiały edukacyjne przeznaczone do wykorzystania na prowadzonych przez nich zajęciach. Bardzo często wielu nauczycieli tworzy materiały na ten sam temat. Połączenie ich kompetencji i tworzonych przez nich treści może zmniejszyć koszty edukacji i poprawić jej jakość. Możliwe jest stworzenie bazy wiedzy dla nauczycieli, która będzie zbiorem obiektów edukacyjnych przeznaczonych do wspólnego wykorzystywania. Ważne jest, by stworzone obiekty edukacyjne spełniały międzynarodowe standardy i były zgodne ze współczesnymi technologiami internetowymi. Stworzenie obiektów edukacyjnych, w których manifesty będą zawierały precyzyjne informacje o treści obiektu pozwoli na poszukiwanie materiałów edukacyjnych.

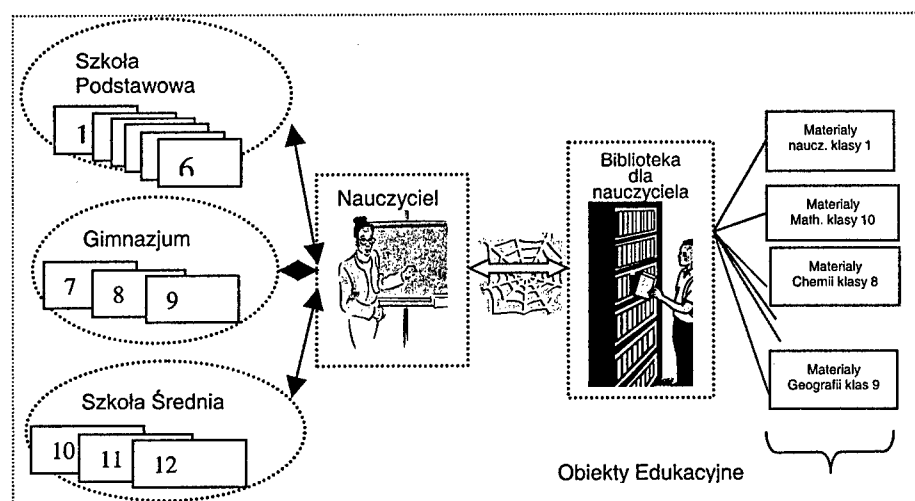
Wykorzystanie Siewerów UDDI i współczesne technologie usług sieciowych umożliwią stworzenie rozproszonych systemów komputerowych wydajnych i przydatnych.

Nauczyciel może tworzyć materiały edukacyjne z wykorzystaniem obiektów już istniejących. Takie rozwiązania zmniejszają koszty tworzenia materiałów edukacyjnych i mogą przyczynić się do poprawy jakości procesu nauczania.

W wielu krajach świata występują problemy w dziedzinie edukacji związane z brakiem nauczycieli lub brakiem środków na prowadzenie działalności edukacyjnej. Wykorzystanie współczesnych technologii usług internetowych może przyczynić się do rozwiązania problemów edukacyjnych w tych krajach.

Na Wydziale Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej wprowadzone są prace, których celem jest stworzenie bazy wiedzy zapęłnianej obiektami edukacyjnymi. System będzie wykorzystywał serwery UDDI i współczesne technologie usług sieciowych. Planuje się, że zostanie wdrożone w państwie Mali (Afryka) przy współpracy malijskiego Ministerstwa Edukacji. Minister Edukacji posiada program rządowy, którego celem jest zmniejszenie kosztów tworzenia i powielanie materiałów edukacyjnych w skali kraju.

Na rysunku 3 przedstawiono koncepcję systemu edukacji SEM (*System of Education in Mali*) [2].



Rys 3. System edukacji SEM

Zakłada się, że wykorzystanie nowoczesnych technologii umożliwia stworzenie systemu, w którym nauczyciele zatrudnieni w kraju będą tworzyć materiały edukacyjne przeznaczone do wykorzystania przez innych nauczycieli. Daje to możliwość zapęłniania systemu materiałami edukacyjnymi przez samych użytkowników. Do zalet systemu można zaliczyć możliwość dowolnej modyfikacji materiałów, łatwe zarządzanie wersjami obiektów edukacyjnych, ułatwienie tworzenia dowolnych materiałów dydaktycznych przez każdego nauczyciela. System daje możliwości administrowania obiektami w skali kraju oraz pozwala na rozwiązanie problemu opłaty za utworzenie materiałów dydaktycznych. Kontrolę nad jakością materiałów, prawami autorskimi i kosztami wytwarzania będzie sprawować ministerstwo. Celem projektu jest podniesienie wskaźnika dostępności do edukacji na wszystkich poziomach nauczania w Mali.

5. ZAKOŃCZENIE

Współczesne technologie usług sieciowe oparte o serwery UDDI pozwalają na poszukiwanie gromadzenie i współdzielenie obiektów różnego typu w sieci Internet. Stwarzają one możliwość budowy bazy wiedzy obiektów edukacyjnych przeznaczonych dla uczniów, jak i dla nauczycieli. Uczącym się pozwala na poszukiwanie materiałów z wykorzystaniem agentów programowych w zależności od poziomu wiedzy i zakładanego profilu kształcenia się. Nauczycielom umożliwia tworzenie materiałów z wykorzystaniem już istniejących obiektów edukacyjnych tworzonych przez innych nauczycieli. Należy przypuszczać, że usługi internetowe oparte o serwery UDDI staną się w przyszłości podstawowym sposobem wykorzystywania zasobów Internetu.

BIBLIOGRAFIA

- [1] Mc Cormack C.: *Web-based Education System*, Wiley 1997.
- [2] Diarra M, Kaczmarek J.: *Budowa repozytoriów materiałów dydaktycznych dla nauczycieli*, Zeszyty Naukowe WEA Politechniki Gdańskiej, nr 19, str. 35-40, 2003.
- [3] Graham S.: *Java. Usługi WWW*, Helion, 2003.
- [4] Bruner R.: *Java w komercyjnych usługach sieciowych*, Helion, 2003.
- [5] Gwoździńska A., Kaczmarek J., Diarra M.: *Poziomy komputeryzacji procesu dydaktycznego*, Zeszyty Naukowe WEA Politechniki Gdańskiej, nr 18, str.18-76, 2002.

THE USE OF UDDI SERVERS FOR EDUCATIONAL SYSTEMS

Summary

Internet contains a meaningful amount of learning materials that are a uniform entity in a form of complete lessons or lectures. The current trend in the e-learning field aims at breaking down learning materials in smaller pieces named learning objects. The paper presents possibilities of UDDI (*Universal Description, Discovery and Integration*) server-based WebServices to store learning objects. We take advantages of XML language, SOAP protocol (*Simple Object Access Protocol*) and WSDL (*WebService Description Language*) standards to describe those possibilities. Such a solution makes possible the creating of both student and teacher's specific knowledge databases that allow teachers to search, make change and use existing multimedia materials. The article shows what educational process organization opportunity at higher level of quality this fact provides.

Jerzy Kaczmarek, Michał Wróbel

Katedra Inżynierii Oprogramowania, Politechnika Gdańska

OBSZARY ZASTOSOWAŃ DYSTRYBUCJI CDLINUX.PL

Streszczenie

System operacyjny GNU/Linux jest używany coraz powszechniej, również jako oprogramowanie stacji roboczych. W 2003 roku powstał projekt cdlinux.pl mający na celu ułatwienie poznawania systemu Linux przez polskich, początkujących użytkowników. W artykule przedstawiono zidentyfikowane wymagania użytkowników, na podstawie których stworzono dystrybucję systemu operacyjnego GNU/Linux. Wyszczególniono obszary potencjalnych zastosowań dystrybucji cdlinux.pl oraz perspektywy dalszego rozwoju.

1. WSTĘP

Stworzony na początku lat dziewięćdziesiątych system operacyjny Linux zyskuje obecnie dużą popularność. Coraz więcej firm i osób prywatnych zwraca się w stronę otwartego oprogramowania. Systemy komputerowe oparte na systemie GNU/Linux zyskały opinię wydajnych, stabilnych i bezpiecznych. Dotychczas Linux i otwarte oprogramowanie było używane głównie w rozwiązaniach serwerowych. Aktualnie można zaobserwować rosnące zainteresowanie wykorzystaniem Linuksa jako systemu operacyjnego przeznaczonego dla stacji roboczych.

Największą przeszkodą w rozprzestrzenianiu się systemu operacyjnego Linux jest powszechna opinia głosząca, że jest to system przeznaczony tylko dla wąskiego grona profesjonalistów. W 2003 roku rozpoczęliśmy projekt o nazwie cdlinux.pl, którego celem jest stworzenie dystrybucji systemu GNU/Linux przeznaczonej dla początkującego użytkownika, dzięki której będzie on mógł zapoznać się z podstawami obsługi systemu oraz z częścią otwartego oprogramowania.

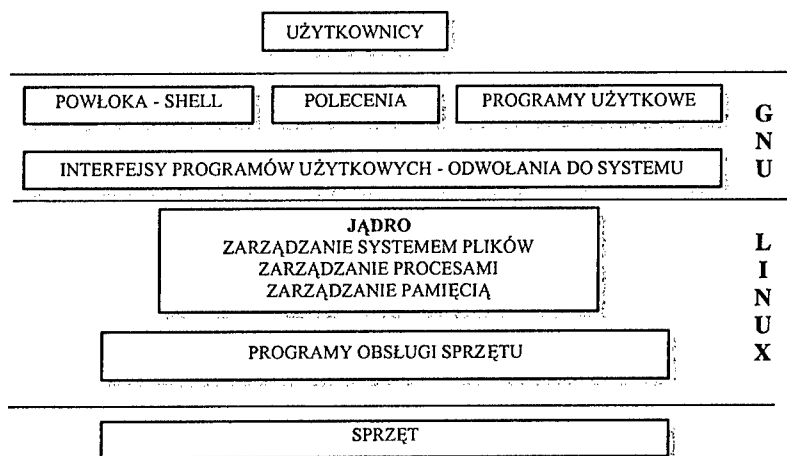
W niniejszym artykule zostaną przedstawione informacje o procesie projektowania dystrybucji cdlinux.pl oraz potencjalnym jej zastosowaniu.

2. DYSTRYBUCJE SYSTEMU OPERACYJNEGO LINUX

Nazwą Linux określane jest jądro systemu operacyjnego. Jest ono tworzone przez tysiące programistów z całego świata, pod kierownictwem jego twórcy – Linusa Torvaldsa, zgodnie z filozofią otwartego oprogramowania (ang. *free software*).

Ideę otwartego oprogramowania zapoczątkował Richard Stallman, który w 1984 roku stworzył projekt GNU (*GNU Not Unix*). Głównym celem było stworzenie systemu operacyjnego, który każdy użytkownik mógłby za darmo używać, redystrybuować i dokonywać dowolnych modyfikacji. Całe oprogramowanie stworzone w ramach projektu jest wydawane na licencji GPL (*GNU General Public Licence*), której celem jest zapewnienie użytkownikowi swobody udostępniania i zmieniania oprogramowania, a więc zagwarantowanie, iż oprogramowanie jest dostępne dla wszystkich użytkowników bez ograniczeń licencyjnych.

Samego jądra systemu operacyjnego w połączeniu ze sprzętem nie można praktycznie wykorzystywać. Jądro stanowi tylko pomost pomiędzy użytkownikiem, programami systemowymi i sprzętem. Na rys. 1 przedstawiono elementy systemu operacyjnego, wraz z obszarami, którymi zarządza. Na całość systemu operacyjnego składają się, oprócz jądra, również systemy plików i powłoka [1]. Dla efektywnego wykorzystywania systemu komputerowego konieczne są dodatkowe programy systemowe i aplikacje użytkowe. Kiedy powstało jądro Linuksa dostępnych już było wiele takich programów, które zostały stworzone w ramach projektu GNU. Najważniejsze z nich to: powłoka (bash), kompilatory (gcc), edytory (emacs) i wiele innych [2]. Po połączeniu jądra Linuksa z oprogramowaniem GNU powstał pełnowartościowy system operacyjny określany mianem GNU/Linux.



Rys.1. Schemat budowy systemu komputerowego

System operacyjny GNU/Linux jest rozprowadzany w formie tak zwanych dystrybucji. Dystrybucją nazywany jest zbiór programów złożony z jądra systemu operacyjnego oraz aplikacji wybranych przez twórców dystrybucji. Obecnie dostępnych jest wiele dystrybucji. Najbardziej popularne to: Mandrake, Red Hat – Fedora Core, Debian, SUSE oraz Slackware. Poszczególne dystrybucje różnią się pomiędzy sobą konstrukcją pakietów z oprogramowaniem, narzędziami do konfiguracji i administracji systemem oraz programami instalacyjnymi. Pozostałe różnice są nieistotne z punktu widzenia użytkownika. Największą dystrybucją jest Debian GNU/Linux [3]. Jest on rozwijany zgodnie z ideą otwartego oprogramowania, tzn. tworzą ją za darmo programiści z całego świata. Do tej dystrybucji włączane jest tylko otwarte oprogramowanie. Dystrybucja cdlinux.pl została stworzona

właśnie na bazie dystrybucji Debian GNU/Linux. Większość oprogramowania jest instalowana z pakietów przygotowanych przez zespół Debiana.

Nowym rodzajem dystrybucji, która zyskuje dużą popularność jest tzw. dystrybucja LiveCD. Dystrybucje tego typu są uruchamiane bezpośrednio z płyty CD-ROM i nie wymagają instalacji na dysku twardym. Pierwszą taką dystrybucją, która zyskała dużą popularność była francuska dystrybucja o nazwie DemoLinux. Obecnie najpopularniejszą dystrybucją tego typu w Europie jest produkt o nazwie Knoppix.

3. ZAŁOŻENIA PROJEKTOWE DYSTRYBUCJI CDLINUX.PL

Głównym celem projektu cdlinux.pl jest stworzenie polskiej dystrybucji systemu operacyjnego GNU/Linux typu LiveCD. Na początku projektu dokonano dokładnej identyfikacji potencjalnych użytkowników dystrybucji. Na jej podstawie zostały wyspecyfikowane wymagania funkcjonalne.

Projekt cdlinux.pl jest przeznaczony przede wszystkim dla użytkowników, którzy do-tychczas nie mieli styczności z systemem GNU/Linux. Użytkownika takiego wyróżnia:

- niechęć do posługiwania się linia poleceń (shellem),
- brak wiedzy w zakresie konfiguracji systemu,
- używanie podstawowego zestawu oprogramowania użytkowego:
 - przeglądarka stron WWW,
 - klient poczty elektronicznej,
 - komunikator internetowy,
 - pakiet biurowy (edytor tekstu, arkusz kalkulacyjny, edytor prezentacji),
 - programy do obsługi multimediiów (muzyka, filmy),
 - programy graficzne.

3.1. Identyfikacja potencjalnych zastosowań

Dystrybucja LiveCD, taka jak cdlinux.pl, może być również używana w innych zastosowaniach, m.in. w edukacji, w tworzeniu sieci komputerowych z bezdyskowymi stacjami roboczymi oraz w prostej instalacji skonfigurowanej dystrybucji GNU/Linux na dysku twardym.

3.1.1. Edukacja

W związku z prowadzonymi na Wydziale Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej zajęć z Systemów Operacyjnych postanowiono zaprojektować dystrybucję cdlinux.pl tak, aby mogła być użyteczna również dla studentów uczęszczających na te zajęcia. Po zbadaniu wymagań studentów okazało się, że są one zasadniczo zgodne z wymaganiami początkującego użytkownika. Konieczne było tylko dodanie następującego oprogramowania:

- interpretery języków skryptowych (bash, perl, python),
- kompilatory (gcc),
- debugger,
- profesjonalne edytory tekstowe (emacs, vim).

3.1.2. Stacje bezdyskowe

Naturalnym wykorzystaniem dystrybucji LiveCD wydawały się bezdyskowe stacje robocze. Terminale tego typu mogłyby zastąpić część komputerów w firmach sektora MSP

(Małe i Średnie Przedsiębiorstwa), co wpłynęłoby na redukcję wydatków, zarówno na sprzęt, jak i administrację. W związku z tym postanowiliśmy dodać oprogramowanie wymagane w tego typu rozwiązaniach, takie jak klienci serwerów plików (NFS, Samba) oraz program do zdalnego łączenia się graficznie z serwerami (VNC). Wykorzystanie komputerów jako bezdyskowych stacji roboczych jest zgodne ze współczesną tendencją do budowy architektury typu „cienki klient”.

3.1.3. System instalacyjny

Dystrybucje LiveCD posiadają pewne wady uniemożliwiające wykorzystanie ich jako stacji roboczych, zwłaszcza w zastosowaniach domowych. Najważniejszą z nich jest brak możliwości instalacji dodatkowego oprogramowania. Jest to znaczne utrudnienie dla użytkownika, który już zapoznał się z dystrybucją `cdlinux.pl` i zgłębił podstawy użytkowania i konfiguracji systemu GNU/Linux. Dlatego postanowiono uzupełnić dystrybucję o program do instalowania skonfigurowanej dystrybucji na dysku twardym.

3.2 Wymagania funkcjonalne

Przedstawiona powyżej identyfikacja potencjalnych użytkowników i zastosowań dystrybucji `cdlinux.pl` pozwoliła stworzyć listę najważniejszych wymagań funkcjonalnych. Określono następujące cechy, którymi musi charakteryzować się dystrybucja:

- automatyczna konfiguracja systemu,
- ograniczenie ingerencji użytkownika w proces konfiguracji,
- obsługa niezbędnych operacji z poziomu trybu graficznego,
- intuicyjny interfejs graficzny.

4. CHARAKTERYSTYKA DYSTRYBUCJI CDLINUX.PL

Pierwsza wersja dystrybucji `cdlinux.pl` została udostępniona w czerwcu 2003 roku. Publiczna prezentacja odbyła się podczas I Bałtyckiego Festiwalu Nauki w Gdańsku. Od tamtej pory ukazało się osiem wersji zawierających poprawki oraz nowe funkcje. Dystrybucja cieszy się powodzeniem – przez pierwsze dwa miesiące 2004 roku została pobrana ponad tysiąc razy z serwera, znajdującego w Centrum Informatycznym Trójmiejskiej Akademickiej Sieci Komputerowej.

Poprzez uruchomione forum i system zgłaszania błędów na stronie <http://www.cdlinux.pl> nawiązano kontakt z użytkownikami. Na forum zostało utworzonych przeszło 130 tematów, na które zostało wysłane ponad 500 odpowiedzi. Dotyczyły one głównie opinii na temat dystrybucji, propozycji dalszego rozwoju i próśb o pomoc. Część użytkowników włączyła się na stałe do projektu testując wstępne wersje oprogramowania.

4.1. Oprogramowanie wykonane w ramach projektu

Zgodnie z wymaganiami funkcjonalnymi konieczne było stworzenie oprogramowania, które w prosty sposób wspomaga użytkownika w konfiguracji i administracji systemem. Stworzony został zbiór oprogramowania określany nazwą *cdlcenter*.

Najważniejszym programem w tym pakiecie jest program konfigurujący system po starcie komputera. Dokonywana jest automatyczna detekcja dostępnych urządzeń, m.in. kart sieciowych, graficznych, muzycznych, dysków twardych i wielu innych. Poprzez proste pytania do użytkownika instalowana jest sieć komputerowa i system X-window.

Na podstawie zidentyfikowanych wymagań użytkownika zostały zaimplementowane programy wspomagające pracę, takie jak:

- program do montowania dysków – wykonuje operacje montowania dysków twardych, płyt CD-ROM i dyskietek w trybie graficznym poprzez kliknięcie odpowiednich ikon,
- program do zarządzania pamięcią swap – pozwala włączać i wyłączać pamięć swap, a także tworzyć pliki swap na partycjach dysku twardego,
- program instalacyjny – pozwala zainstalować dystrybucję na dysku twardym, dzięki czemu można doinstalowywać dowolne oprogramowanie,
- program do tworzenia własnych wersji dystrybucji cdlinux.pl (aktualnie w fazie rozwojowej).

4.2. Rodzaje dystrybucji

Aktualnie dostępne są dwa rodzaje dystrybucji cdlinux.pl, tzw. *mały* i *duży*. Różnią się one wielkością zajmowaną na płycie CD-ROM oraz ilością zawartego oprogramowania.

Dystrybucja *cdlinux.pl mały* zajmuje tylko 200 MB i mieści się na małej płycie CD-ROM o rozmiarze 9 cm. W tym rozwiązaniu istnieje również możliwość umieszczenia całego oprogramowania w pamięci RAM, co pozwala na wyjęcie płyty z dystrybucją z napędu. Dzięki temu możliwe jest oglądanie filmów DVD lub słuchanie muzyki z innych płyt CD. Wersja *mały* zawiera podstawowe oprogramowanie przeznaczone dla początkującego użytkownika: przeglądarkę stron WWW, klienta poczty, komunikator internetowy, programy do odtwarzania muzyki i filmów, przeglądarkę plików graficznych oraz edytor tekstu.

Drugi rodzaj dystrybucji to tzw. *cdlinux.pl duży*. Zawiera on kompletne oprogramowanie dołączane zwykle do dystrybucji przeznaczonych do zastosowań domowych i biurowych. Zajmuje na płycie około 650 MB, jednak dzięki zastosowanej kompresji umieszczone jest na nim prawie dwa razy więcej danych. Oprócz oprogramowania zawartego w *cdlinux.pl mały* znajduje się na nim pakiet biurowy OpenOffice.org z arkuszem kalkulacyjnym i programem do prezentacji oraz programy z pakietu KDE.

5. ZAKOŃCZENIE

Dystrybucja cdlinux.pl została zaprojektowana i stworzona z myślą o początkującym użytkowniku. W związku z rosnącym zainteresowaniem Linuksem może stać się pomocnym narzędziem w nauce obsługi i administracji systemem Linux.

Dalszy rozwój dystrybucji cdlinux.pl będzie zmierzał w kierunku stworzenia instalatora, który umożliwi prostą instalację systemu operacyjnego opartego na Debianie na dysku twardym. Planuje się stworzenie narzędzia do samodzielnej rozbudowy dystrybucji cdlinux.pl o wybrane przez użytkownika oprogramowanie.

Projekt jest zgodny z najnowszymi kierunkami rozwoju informatyki, według których wszystkie operacje powinny być wykonywane na dedykowanych serwerach, a nawet w rozproszonej sieci serwerów (*Grid Computing*). Stacje robocze klientów byłyby wyposażone tylko w programy tzw. cienkich klientów, poprzez które łączyłyby się z serwerami.

Zgodnie ze współczesnymi tendencjami wydaje się, że dystrybucje systemu Linux typu LiveCD, które organizują pracę komputerów w architekturze typu „cienki klient” będą coraz popularniejsze.

BIBLIOGRAFIA

- [1] Silberschatz A., Galvin P. B.: *Podstawy systemów operacyjnych*, WNT, Warszawa, 2002.
- [2] Prata S., Martin D.: *Biblia systemu UNIX V*, Warszawa, 1994
- [3] Camou M., Goerzen J.: *Debian Linux*, Helion, Gliwice 2001

APPLIANCE AREA OF CDLINUX.PL DISTRIBUTION**Summary**

The use of GNU/Linux operating system is growing rapidly, also as a software for workstations. In 2003 cdlinux.pl project was launched. Its aim was to create easy for use GNU/Linux distribution designed for Polish beginners. In the article there are presented identified users requirements based on which cdlinux.pl distribution was created. There is specified area of potential appliance of cdlinux.pl distribution and perspective of further development.

Radosław P. Katarzyniak

Instytut Sterowania i Techniki Systemów, Politechnika Wrocławska

**ZASTOSOWANIE ALGORYTMU HEURYSTYCZNEGO DO
WYZNACZANIA KLAS UŻYTKOWNIKÓW KOOPERUJĄCYCH
W SIECIOWYM SYSTEMIE INFORMATYCZNYM**

Streszczenie

Przedstawiono model analizy zachowań użytkowników sieciowego systemu teleinformatycznego. Skoncentrowano uwagę na automatycznym wyznaczaniu grup użytkowników tworzących się spontanicznie podczas korzystania z systemu teleinformatycznego. Zdefiniowano zadanie wyznaczania modelu kooperacji i pokazano, w jakim sposób zadanie to można sprowadzić do NP.-zupełnego zadania wyznaczania profilu zbioru podziałów. Wskazano algorytm heurystyczny wyznaczający przybliżony model klas kooperujących użytkowników systemu.

1. WSTĘP

Analiza zachowań użytkowników sieciowych systemów teleinformatycznych jest ważnym elementem ustalania wpływu, który sieci teleinformatyczne wywierają na funkcjonowanie współczesnego społeczeństwa informacyjnego. Jednym z aspektów zachowań społecznych badanych w tym zakresie z powodów ekonomicznych i socjologicznych jest tworzenie się wśród użytkowników grup osób powiązanych ze sobą intensywną wymianą informacji. Występowanie takiej wymiany może wynikać z różnych przyczyn np. członkowie grup mogą posiadać wspólny cel lub podobne zainteresowania zawodowe. Wiedza o tego typu związkach powinna mieć istotny wpływ na zarządzanie użytkownikami systemu.

Wyznaczanie samodzielnie tworzących się i rozwijających grup użytkowników sieciowego systemu teleinformatycznego jest zadaniem trudnym. Wiedza o zaistnieniu pojedynczej transakcji pomiędzy użytkownikami nie jest bowiem w tym przypadku wystarczająca do wyznaczenia faktycznie istniejącego podziału użytkowników na grupy kooperujących osób, ponieważ o istnieniu grupy można wnioskować dopiero z całokształtu obserwacji działającego systemu oraz kompletnej listy transakcji zrealizowanych w nim pomiędzy użytkownikami. Transakcje definiujące kooperującą grupę cechować musi bowiem powtarzalność.

W pracy zadanie wyznaczenia klas kooperujących użytkowników rozważane jest dla sieciowego systemu teleinformatycznego, w którym kooperacja użytkowników badana jest

cyklicznie dla zadanego przedziału czasowego stanowiącego np. godziny pracy przedsiębiorstwa korzystającego z sieciowego systemu teleinformatycznego. Założono dalej, że sieciowy system informatyczny poddawany jest obserwacjom w ww. przedziale czasowym w celu ustalenia listy transakcji zrealizowanych w tym właśnie przedziale pomiędzy poszczególnymi użytkownikami. Na podstawie statystyki zrealizowanych transakcji wyznaczana jest lista klas (podział) użytkowników, które w sposób naturalny wyróżniły się w zadanym przedziale czasowym. Natomiast właściwy model podziału użytkowników ustalany jest na podstawie listy podziałów użytkowników wyznaczonych dla zadanego przedziału czasowego rozpatrywanych np. dla każdego dnia roboczego konkretnego miesiąca kalendarzowego.

Zakłada się dalej, że realizacja postawionego wyżej zadania wyznaczenia podziału użytkowników przeprowadzana jest w oparciu o relatywnie prosty model teoretyczny. W modelu tym przyjmuje się, że zadanie wyznaczenia klas użytkowników sprowadzić można do rozwiązania NP.-zupełnego problemu wyznaczenia reprezentanta zbioru podziałów. Pokazano, w jaki sposób do rozwiązania tego zadania zastosować algorytm heurystyczny podany w pracy [1].

2. ZACHOWANIA UŻYTKOWNIKÓW SYSTEMU

Niech dany będzie sieciowy system teleinformatyczny przedsiębiorstwa, z którego korzystają użytkownicy $U = \{u_1, \dots, u_N\}$. Zachowania użytkowników systemu badane są dla serii $\langle p^{(1)}, p^{(2)}, \dots, p^{(K)} \rangle$ regularnie powtarzających się przedziałów czasowych $p^{(i)}$, $i=1, 2, \dots, K$. Przykładowo, mogą to być godziny pracy [08.00 - 17.00] w przedsiębiorstwie rozpatrywane w każdy roboczy dzień wybranego miesiąca kalendarzowego. Rezultatem obserwacji zachowań użytkowników mających miejsce w zadanym konkretnym przedziale czasowym jest statystyka zrealizowanych w nim transakcji:

Określenie 2.1. Zachowania użytkowników sieciowego systemu teleinformatycznego w przedziale czasowym $p^{(i)}$, $i=1, 2, \dots, K$, dane są statystyką zrealizowanych w nim transakcji. zachowania te opisane są macierzą kwadratową

$$S^{(i)} = [s_{p,q}^{(i)}]_{N \times N} \quad (2.1)$$

w której $s_{p,q}^{(i)} \geq 0$ jest liczbą zaobserwowanych transakcji zrealizowanych w przedziale $p^{(i)}$ pomiędzy użytkownikami u_p oraz u_q .

Przyjmuje się dalej, że wstępna statystyka opisująca intensywność współpracy użytkowników systemu teleinformatycznego w przedziale $p^{(i)}$, $i=1, 2, \dots, K$, poddawana jest normalizacji. Znormalizowanie macierzy pozwala porównać względną intensywność kooperacji użytkowników wyznaczoną dla jednego przedziału z analogiczną intensywnością charakteryzującą współpracę tych samych użytkowników w innym przedziale czasowym:

Określenie 2.2. Znormalizowana macierz zachowań użytkowników w przedziale czasowym $p^{(i)}$, $i=1, 2, \dots, K$ dana jest jako:

$$N^{(i)} = [n_{p,q}^{(i)}]_{N \times N} = \left(\frac{1}{\text{Max}^{(i)}} \right) \cdot s_{p,q}^{(i)} \quad (2.2)$$

gdzie $\text{Max}^{(i)} = \max_{p,q \in \{1, 2, \dots, N\}} (s_{p,q}^{(i)})$, dla $i=1, 2, \dots, K$.

W proponowanym modelu wyznaczania podziału użytkowników na klasy kooperujących osób przyjmuje się dalej, że o zachodzeniu kooperacji dwóch użytkowników u_p oraz p_q w przedziale czasowym $p^{(i)}$, $i=1,2,\dots,K$ można mówić, gdy znormalizowana liczność transakcji zrealizowanych pomiędzy nimi osiąga zadany poziom krytyczny $\lambda \in (0,1]$. Wartość progu λ ustalana musi być w sposób empiryczny dla podanego sieciowego systemu teleinformatycznego oraz z uwzględnieniem specyfiki wykorzystującej go grupy użytkowników. Ustalanie wspomnianej specyfiki wykracza poza zakres zagadnień ściśle obliczeniowych. Uwzględnienie progu λ prowadzi do następującego przekształcenia macierzy $N^{(i)}$ w macierz binarną $B^{(i)}$:

Określenie 2.3. λ -poziomowa macierz zachowań użytkowników w przedziale czasowym $p^{(i)}$, $i=1,2,\dots,K$, dana jest jako

$$B^{(i)} = [b_{p,q}^{(i)}]_{N \times N} \quad (2.3)$$

$$\text{w której } b_{p,q}^{(i)} = \begin{cases} 1 & \text{gdy } n_{p,q}^{(i)} \geq \lambda \\ 0 & \text{gdy } n_{p,q}^{(i)} < \lambda \end{cases}, \text{ dla } i=1,2,\dots,K.$$

Macierz $B^{(i)}$ reprezentuje graf skierowany, który opisuje stan kooperacji zrealizowanej przez użytkowników U w przedziale czasowym $p^{(i)}$, $i=1,2,\dots,K$. Z socjologicznego punktu widzenia racjonalne jest przyjąć, że grupę kooperujących użytkowników tworzą osoby, które w sposób bezpośredni lub pośredni związane były zrealizowanymi transakcjami. W szczególności, jeżeli z macierzy $B^{(i)}$ wynika, że osoba u_m kooperowała z osobą u_n (tj. $b_{m,n}^{(i)}=1$) i osoba u_m kooperowała także z osobą u_o (tj. $b_{m,o}^{(i)}=1$), to w sensie społecznym przy aktywnym i autonomicznym udziale osoby u_m wykreowana została grupa trzech kooperujących użytkowników $\{u_m, u_n, u_o\}$. W modelu podejście to uogólnione zostaje w następujący sposób:

Określenie 2.4. Niech dana będzie λ -poziomowa macierz $B^{(i)}$ zachowań użytkowników w przedziale czasowym $p^{(i)}$, $i=1,2,\dots,K$. Przyjmuje się, że binarna relacja kooperacji użytkowników $K^{(i)} \subseteq U \times U$ dana jest rekurencyjnie w następujący sposób:

$$\langle u, u \rangle \in K^{(i)} \quad (2.4)$$

$$\text{jeżeli } b_{m,n}^{(i)}=1 \text{ lub } b_{n,m}^{(i)}=1, \text{ to } \langle u_m, u_n \rangle \in K^{(i)} \quad (2.5)$$

$$\text{jeżeli } \langle u_m, u_n \rangle \in K^{(i)} \text{ oraz } (b_{n,o}^{(i)}=1 \text{ lub } b_{o,n}^{(i)}=1), \text{ to } \langle u_m, u_o \rangle \in K^{(i)} \quad (2.6)$$

Łatwo zauważyć, że relacja $K^{(i)}$ jest relacją zwrotną, symetryczną i przechodni a, czyli jest relacją równoważności i wyznacza podział $P^{(i)}$ zbioru U na klasy użytkowników, którzy w przedziale $p^{(i)}$ kooperowali uczestnicząc we wspólnym "łańcuchu" transakcji.

Wprowadzone określenia prowadzą do sytuacji, w której zachowania użytkowników w przedziałach czasowych $p^{(i)}$, $i=1,2,\dots,K$ opisane są odpowiednio podziałami $P^{(i)}$, $i=1,2,\dots,K$.

3. MODEL KOOPERACJI

Zbiór podziałów $P = \{P^{(i)}; i=1,2,\dots,K\}$ interpretować należy jako wstępnie przetworzone dane empiryczne pochodzące z obserwacji użytkowników rzeczywistego sieciowego systemu teleinformatycznego. Zbioru P nie można jednak interpretować jako modelu kooperacji specyfikującego trwale występującą tendencję do współpracy użytkowników.

Zachowania użytkowników obserwowane w poszczególnych przedziałach $P^{(i)}$ nie muszą być bowiem identyczne i dopiero analiza tendencji obserwowanej dla serii przedziałów czasowych $P^{(i)}$ dostarczyć może dokładniejszego wglądu w naturalnie ukształtowany podział użytkowników. Poniżej pokazane jest, w jaki sposób wyznaczenie modelu kooperacji ze zbioru P sprowadzić można do rozwiązania zadania wyznaczania profilu zbioru podziałów.

3.1. Zadanie wyznaczenia modelu kooperacji

Przyjmijmy następujące określenie:

Określenie 3.1.1. Niech dane będą zbiór użytkowników $U=\{u_1, \dots, u_N\}$, zbiór $E(U)$ wszystkich podziałów zbioru U oraz zbiór podziałów $P=\{P^{(i)}: i=1,2,\dots,K\}$ określony powyżej. Model kooperacji użytkowników, wyznaczony na podstawie serii obserwacji przeprowadzonych dla przedziałów czasowych $P^{(i)}$, $i=1,2,\dots,K$, dany jest jako podział $P^* \in E(U)$ spełniający warunek Mediany Kemeny'ego:

$$\sum_{i=1}^K d(P^*, P^{(i)}) = \min_{R \in E(U)} \left(\sum_{i=1}^K d(R, P^{(i)}) \right) \quad (3.1.1)$$

gdzie funkcja o sygnaturze $d: E(U) \times E(U) \rightarrow N \cup \{0\}$ jest funkcja odległości zdefiniowaną dla przestrzeni wszystkich podziałów zbioru U .

Zaproponowane określenie modelu kooperacji użytkowników wyłonionego z danych empirycznych dotyczących transakcji zachodzących pomiędzy użytkownikami oznacza, że w proponowanym ujęciu zadanie wyznaczenia modelu sprowadzone zostaje do zadania wyznaczenia profilu zbioru podziałów [2]. Istnieją liczne metody mierzenia odległości między podziałami [3]. Jedną z najprostszych zastosowano w pracy [1]:

$$d(P, Q) = \left(\frac{1}{2} \right) \cdot \sum_{i=1}^N \sum_{j=1}^N |p_{ij} - q_{ij}| \quad (3.1.2)$$

gdzie $P, Q \in E(U)$. Dla tak zdefiniowanej odległości zadanie wyznaczenia P^* należy do klasy zadań NP.-zupełnych [3]. Stąd należy przyjąć, że nie istnieje efektywna procedura wyznaczająca model P^* i w konsekwencji konieczne jest zaproponowanie heurystycznej metody obliczeniowej wyznaczającej przybliżone rozwiązanie P^H . W proponowanym podejściu do wyznaczenia P^H wykorzystana zostaje heurystyka podana w pracy [1].

3.2. Algorytm heurystyczny

Algorytmy heurystyczne do wyznaczania profilu zbioru podziałów przedstawione w pracy [1] odwołują się do pojęcia iloczynu podziałów, maszynowej reprezentacji podziału w postaci listy klas oraz kilku działań na reprezentacji maszynowej. Pojęcia te definiowane są w następujący sposób:

Określenie 3.2.1. Niech dane będą podziały $T, Q, R \in E(U)$ oraz odpowiadające im relacje binarne $\tilde{t}, \tilde{q}, \tilde{r} \subseteq U \times U$. Podział T jest iloczynem podziałów Q i R (formalnie $T=Q \otimes R$) wtedy i tylko wtedy, gdy zachodzi równość $\tilde{t} = \tilde{q} \cap \tilde{r}$.

Określenie 3.2.2. Niech dany będzie podział $T \in E(O)$ niepustego zbioru obiektów O .

Przyjmuje się, że:

- a) symbol $\text{lenght}(T)$ oznacza liczbę bloków podziału T .
- b) lista $T = [T_1, T_2, \dots, T_{\text{lenght}(T)}]$ jest maszynową reprezentacją podziału T , gdzie T_i są poszczególnymi blokami podziału. $T[i]$ oznacza i -ty element listy.
- c) działanie $\text{erase}(T, Z)$, gdzie $Z \subseteq O$, $Z \neq \emptyset$, podstawia $T[i] = \emptyset$ wtedy i tylko wtedy, gdy $T[i] = Z$.
- d) działanie $\text{pack}(T)$ usuwa z listy T wszystkie $T[i] = \emptyset$.

Przebieg algorytmu heurystycznego jest następujący:

```

Dane wejściowe:    $P = \{P^1, P^2, \dots, P^K\}$ ,
                   $W$  – wartość progu włączenia klasy
Dane wyjściowe:    $P^H$  – rozwiązanie heurystyczne
begin
   $\Omega := P^1 \otimes P^2 \otimes \dots \otimes P^K$ ;
   $\Omega = [E_1, E_2, \dots, E_{\text{lenght}(\Omega)}]$ 
   $P^H := \emptyset$ ;
  While  $\text{lenght}(\Omega) \neq 0$  then begin
    HClass =  $\Omega[1]$ ;
     $\Gamma = \Omega[1]$ ;
     $\text{erase}(\Omega, \Omega[1])$ ;
    for  $j := 2$  to  $\text{lenght}(\Omega)$  do begin
       $s := 0$ ;
      for  $k := 1$  to  $K$  do if  $(\exists Z \in P^k. \Omega[j] \cup \Gamma \subseteq Z)$  then  $s := s + 1$ ;
      if  $s \geq W$  then begin
        HClass =  $\text{HClass} \cup \Omega[j]$ ;
         $\text{erase}(\Omega, \Omega[j])$ ;
      end
    end
    end
     $\text{pack}(\Omega)$ ;
     $P^H = P^H \cup \{\text{HClass}\}$ 
  end
end.
```

Własności powyższego algorytmu heurystycznego oraz jakość wyznaczanego rozwiązania podano w [1]. Przykład ilustrujący działanie algorytmu podano w [4].

6. ZAKOŃCZENIE

Analiza zachowań użytkowników sieciowych systemów teleinformatycznych jest ważnym obszarem przetwarzania informacji. Uzyskiwane w niej wyniki posiadają znaczenie praktyczne. Niestety zbudowanie modelu zachowania użytkowników na podstawie danych empirycznych pochodzących z obserwacji rzeczywistych systemów jest zazwyczaj trudne, gdyż uzyskane dane mają charakter przybliżony, są niekompletne i często dotyczą jedynie wybranego fragmentu zachowań użytkowników. Niejednokrotnie dodatkowym problemem staje się sama natura rozwiązywanego zadania analizy danych. Sytuacja

rozpatrywana w pracy dostarcza takiego właśnie przykładu. Analiza zachowań użytkowników okazuje się bowiem zadaniem NP.-zupełnym i wymaga zastosowania przybliżonych metod heurystycznych.

Rozwiązanie teoretyczne zaprezentowane w pracy stanowi zamknięcie wstępnego etapu badań nad wykorzystaniem ogólnej heurystyki wyznaczającej profil zbioru podziałów [1] do analizy konkretnych zachowań użytkowników środowiska teleinformatycznego. W kolejnych etapach przedsięwzięcia przewidziano praktyczne wykorzystanie proponowanego modelu oraz weryfikację uzyskiwanych rezultatów.

BIBLIOGRAFIA

- [1] Katarzyniak R, Nguyen N.T.: *Heuristics for consensus partition problem.*, Advances in Modelling & Analysis, A, Vol. 13, No. 4, pp. 1-12, 1992.
- [2] Daniłowicz C., Nguyen N.T.: *Metody wyboru reprezentacji podziałów i pokryć uporządkowanych.*, Wydawnictwo Politechniki Wrocławskiej, Wrocław 1992.
- [3] Daniłowicz C., Ngoc T.N., Jankowski Ł., *Metody wyboru reprezentacji stanu wiedzy agentów w systemach multiagenkich.* Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2002.
- [4] Katarzyniak R., *Wykorzystanie algorytmu heurystycznego do uzgadniania tezauryśa przez system wieloagentowy.*, W: Bubnicki Z., Grzech A. (red.), *Materiały konferencyjne V Krajowej Konferencji pt. Inżynieria Wiedzy i Systemy Ekspertowe*, Wrocław 11-13 czerwca 2003), str. 262-270, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2003.

APPLYING HEURISTIC TO EXTRACT CLASSES OF USERS COOPERATING IN NETWORKED INFORMATION SYSTEM

Summary

An approach to the analysis of user behavior has been presented. The target was to model the process of automatic extraction of users' groups emerging spontaneously in information system. The problem of cooperation model extraction has been defined and proved to be equivalent to NP.-complete problem of computing consensus partition. A certain heuristic has been chosen and applied in order to solve the problem of extracting the model of users' group.

Monika Koprowska, Rafał Sawzdargo

Instytut Sterowania i Techniki Systemów, Politechnika Wrocławska

ZARZĄDZANIE USŁUGAMI W NOWOCZESNYCH SYSTEMACH WEBOWYCH

Streszczenie

W artykule omówiono zagadnienia związane z tworzeniem i zarządzaniem usługami sieciowymi (*Web Services*). Przedstawiona została koncepcja oraz zastosowania sieci semantycznej (*Semantic Web*), która w założeniach ma ułatwiać zarządzanie wiedzą w systemach internetowych. Omówiono metody definiowania usług sieciowych, a także sposoby nawiązywania komunikacji między aplikacjami oraz systemami informatycznymi, działającymi na różnych platformach programowo-sprzętowych. Podano przykłady obecnie najpopularniejszych technologii związanych z zarządzaniem usługami webowymi (SOAP, CORBA). Sformułowano wnioski na temat efektywności omówionych mechanizmów wykrywania usług webowych.

1. WSTĘP

Odkąd komputery z autonomicznych urządzeń przekształcone zostały w elementy składowe ogólnosiwiatowej struktury, jaką jest World Wide Web (WWW), konieczne stało się zaopatrzenie ich w mechanizmy ułatwiające wymianę danych. Związany z tym rozwój technologii informatycznych doprowadził do powstania wielu systemów, standardów i protokołów. Równocześnie pojawiła się potrzeba powiązania zastosowanych rozwiązań tak, aby możliwa była ich efektywna współpraca.

W związku z tym, kilka lat temu pojawia się koncepcja sieci semantycznej (ang. *Semantic Web*) [1]. W zamyśle jej twórcy ma to być kolejna wersja obecnie działającej sieci, uzupełniona o informację semantyczną. Proces budowy inteligentnych systemów zarządzania informacją nie może obejść się bez możliwości pozyskiwania wiedzy. Jest to jednak utrudnione przez fakt istnienia wielu źródeł oraz różnorodność takiej wiedzy. Stąd nacisk na rozwiązanie problemów związanych ze sposobem reprezentacji wiedzy w systemach internetowych.

Ważnym aspektem będzie również dostęp do danych zamieszczonych w sieci oraz metody ich pozyskiwania, co wiąże się z rozwojem usług sieciowych (*Web Services*). Liczące się na tym polu specyfikacje to *Common Object Request Broker Architecture* (CORBA), opracowana przez Object Management Group, oraz *Simple Object Access Protocol* (SOAP). Zostaną one przedstawione w dalszej części artykułu, wraz z porównaniem ich najważniejszych – z punktu widzenia projektanta – własności.

2. SIEĆ SEMANTYCZNA (SEMANTIC WEB)

Podstawowym zadaniem języka HTML jest prezentacja zawartości stron WWW. Natomiast utrudniona jest ekstrakcja i automatyczne przetwarzanie reprezentowanej w nich wiedzy. Rozwiązanie tych problemów umożliwić ma koncepcja sieci semantycznej [1]. Standardy sieci semantycznej ustanawiają globalną strukturę niezbędną nie tylko do wymiany informacji (jak to ma obecnie miejsce w sieci WWW), ale również metadanych oraz ontologii. Ich wpływ na tworzone w przyszłości dla systemów informacyjnych aplikacje obejmuje różne dziedziny współczesnej nauki, a ich zasięg oddziaływania może być porównywany z tym, jaki obecnie ma sieć WWW.

Architektura sieci semantycznej zakłada wykorzystanie do reprezentacji danych języka XML, który pozwala użytkownikowi definiować własne znaczniki. Zaletą XML jest również format dokumentu umożliwiający automatyczne jego przetwarzanie. Nie umożliwia on jednak określania znaczenia prezentowanych danych. Znaczenie to wprowadza dopiero standard RDF (*Resource Description Framework* [2]), w którym dane reprezentowane są w postaci uporządkowanych trójek {*podmiot, predykat, obiekt*}. Podmiot oraz obiekt mogą być zasobami sieciowymi (określanymi przez uniwersalne identyfikatory zasobów – URI) albo literałami, a predykaty opisują relacje między nimi. RDF pozwala na budowanie twierdzeń dotyczących istniejących wyrażeń oraz definiowanie obiektów z wykorzystaniem struktur hierarchicznych i relacji dziedziczenia (*RDF Schema*). Ontologie, stanowiące kolejną warstwę architektury sieci semantycznej, dostarczają natomiast mechanizmów wnioskowanie na podstawie danych.

Prace nad koncepcją przyszłej sieci są istotną częścią badań prowadzonych przez World Wide Web Consortium (W3C). Ta sama organizacja zajmuje się opracowywaniem standardów wspierających usługi sieciowe. Wśród zastosowań sieci semantycznej, które realizowane będą przez te usługi wyróżnić można między innymi:

1. Inteligentne wyszukiwanie informacji – wprowadzenie sieci semantycznej zmieni wyszukiwanie oparte na słowach kluczowych w wydajniejsze, oparte na znaczeniu informacji, bez potrzeby angażowania w to użytkownika. Przeprowadzone będzie ono przez autonomiczne programy potrafiące wymieniać się zdobytą wiedzą. Zamiast wpisywania listy słów kluczowych użytkownicy będą mieli również możliwość formułowania zapytań na temat interesującej ich informacji w formie zbliżonej do języka naturalnego [3].
2. Handel elektroniczny – obejmuje możliwości automatycznego budowania profili użytkowników, zautomatyzowanie negocjacji dzięki zastosowaniu serwisów aukcyjnych oraz agentów wykorzystujących ontologie do wyszukania produktu, o który pyta klient [4]. Efektywne wyszukiwanieżądanego produktu jest obecnie jednym z najważniejszych problemów powstrzymujących spodziewany wzrost usług związanych z handlem elektronicznym.
3. Zarządzenie przedsiębiorstwami – semantyczny intranet (ang. *semantic intranets*) stanie się medium do komunikacji pomiędzy grupami pracowników, użytkowników oraz inwestorów, a także między organizacjami [5].

3. USŁUGI SIECIOWE (*WEB SERVICES*)

Termin Web Services (WS) oznacza usługi dynamicznie łączące się przez sieć i wymieniające między sobą komunikaty. Zastosowanie w nich XML-a uniezależnia komunikację programów od języka, w którym zostały napisane, platformy sprzętowej oraz oprogramowania wykorzystanego do obsługi komponentów wchodzących w skład usługi.

Mimo że na świecie koncepcja Web Services cieszy się coraz większym zainteresowaniem, architektura tych usług jest nadal niekompletna. Tworzeniem jej zajmuje się Web Services Architecture Working Group (WSAWG), działająca w ramach W3C. Wśród wymagań sprecyzowanych przez grupę roboczą znajduje się założenie o kompatybilności usług sieciowych z powstającą w ramach innych prac W3C siecią semantyczną [6,7].

Również poza W3C trwają badania dotyczące dynamicznego wyszukiwania usług podczas pracy aplikacji oraz możliwości wykorzystania do tego języków skryptowych (np. JavaScript). W tą działalność badawczą wpisuje się m.in. organizacja OASIS (*Organization for the Advancement of Structured Information Standards*) [8]. Ponadto czołowe firmy branży komputerowej oferują darmowe zestawy narzędzi, które pozwalają na proste i efektywne tworzenie WS. Umożliwiają one na przykład przekształcanie istniejących komponentów (takich jak COM czy JavaBeans) w usługi sieciowe. Firma IBM rozwija nieodpłatne narzędzia w kategorii Alphaworks (w planach firmy jest zbudowanie kompletnego środowiska dla usług sieciowych – *Application Framework for Web Services*). Na omawianej technologii usług internetowych opiera się też komercyjna platforma Microsoft .NET.

W przyszłości usługi sieciowe będą prawdopodobnie szeroko stosowane do przekazywania informacji (np. wyników notowań giełdowych), czy zapewnienia pracy transakcyjnej (obsługa giełd, systemy rezerwacji, itp.).

4. TECHNOLOGIE WSPOMAGAJĄCE USŁUGI SIECIOWE

Jak już było wspomniane, możliwość komunikacji między aplikacjami rozproszonymi w sieci wymaga wcześniejszego opracowania standardu formatowania i przesyłania informacji. Jednocześnie zaznaczyć należy, że brak odpowiedniego poziomu bezpieczeństwa usług sieciowych skutecznie powstrzymuje ich intensywny rozwój. Użytkownicy, którzy planują w przyszłości stosować WS mogą jednak skorzystać z jednej z przedstawionych poniżej technologii. W dalszej części artykułu przedstawione zostanie porównanie ich najważniejszych, z punktu widzenia użytkownika, właściwości.

4.1. CORBA (*COMMON OBJECT REQUEST BROKER ARCHITECTURE*)

Architektura technologii CORBA zakłada istnienie standardowego zbioru funkcji, pozwalającego na łączenie się obiektów dostarczających usługi z obiektami korzystającymi z usług [9,10]. Podstawową funkcją CORBY jest zatem umożliwienie klientom korzystanie z usług dostarczanych przez rozproszone obiekty. Usługi te zdefiniowane są w języku IDL (*Interface Definition Language*).

Serwisem, który obsługuje zlecenie kierowane do zdalnego obiektu jest tzw. *Object Request Broker* (ORB). Ma on za zadanie zlokalizowanie obiektu w sieci, dostarczenie mu zlecenia klienta oraz zwrócenie temu ostatniemu wyników. Ponadto mechanizm ten nie zależy od wzajemnej lokalizacji obiektów, co oznacza, że klient postępuje identycznie

zlecając wykonanie usługi tak obiektowi zainstalowanemu w tym samym procesie, jak i obiektowi na innym komputerze w sieci [11,12]. Klient zgłaszając chęć wykonania operacji na zdalnym obiekcie wysyła odpowiednie żądanie do *namiastki* – reprezentanta obiektu CORBY po stronie klienta – a ta angażuje w wykonanie zadania mechanizmy dostarczone przez ORB uruchomiony na lokalnej maszynie. Lokalny ORB za pomocą protokołu komunikacyjnego IIOP (Internet Inter-ORB Protocol) bazującego na standardzie TCP/IP przesyła żądanie do ORB zlokalizowanego na maszynie docelowej. Ten dostarcza zlecenie obiektowi CORBY reprezentowanej przez tzw. szkielet, który wcześniej lokalizuje. Szkielet po dokonaniu translacji otrzymanego wywołania na używany lokalnie format wywołuje odpowiednią metodę z implementacji obiektu. Zwracana przez nią wartość jest transformowana przez szkielet do postaci zgodnej z oczekiwaniami przez klienta i wysyłana do niego poprzez IIOP. Ważną cechą ORB jest niezależność zlecenia od języka, za pomocą którego jest ono implementowane. W efekcie tego klient generujący zlecenia może być napisany w innym języku programowania, niż realizujący je obiekt CORBY. Mechanizm ORB odpowiedzialny jest za dokonanie niezbędnych w tym wypadku translacji.

Częścią standardu CORBY jest definicja całego zbioru usług wspomagających współdziałanie rozproszonych obiektów. Są one znane jako serwisy CORBY, w skrócie COS – *CORBA Object Services*.

Schemat projektowania i implementacji obiektowo zorientowanej aplikacji rozproszonej przy użyciu standardu IDL przedstawiony został poniżej:

1. Zdefiniowanie zdalnych interfejsów IDL – notacja podobna do tej z C++ czy Javy jest łatwa do opanowania dla każdego programisty. Większość powszechnie stosowanych języków programowania (w tym C, Ada czy Java) wspiera translację IDL, daje to możliwość szerokiego wyboru środków jakie możemy użyć w procesie tworzenia aplikacji klient-serwer.
2. Kompilacja zdalnych interfejsów – używamy tutaj kompilatora IDLJ w odniesieniu do plików źródłowych wraz z definicjami zdalnych interfejsów. W procesie kompilacji powstaje wersja interfejsów w Javie, oraz namiastki i szkielety, które pozwolą naszej aplikacji na komunikowanie się poprzez Internet.
3. Implementacja serwera – kod po stronie serwera zawierać musi przede wszystkim implementacje metod deklarowanych w zdalnym interfejsie. Ponadto musi zawierać fragment odpowiedzialny za uruchomienie serwisu ORB oraz oczekiwanie na nadejście ze strony klienta zlecenia wywołania którejś z usług (metod) interfejsu;
4. Implementacja klienta – po stronie klienta aplikacja powinna zawierać namiastki wygenerowane przez kompilator IDLJ w celu uruchomienia ORB, zlokalizowania serwera za pomocą specjalnego serwisu nazw IDL, pobrania referencji obiektu CORBY i wywołania jego metod.
5. Uruchomienie aplikacji – krok ten sprowadza się już tylko do uruchomienia serwisu nazw, wystartowania serwera i na końcu klienta.

4.2. SOAP (SIMPLE OBJECT ACCESS PROTOCOL)

Innym rozwiązaniem problemu współpracy różnych technologii jest SOAP. Umożliwia on wywołanie funkcji oraz korzystanie z obiektów udostępnianych przez serwery sieciowe. Komunikacja pomiędzy klientem a serwerem odbywa się za pomocą języka XML. SOAP definiuje gramatykę XML służącą do określania nazw metod, typów

parametrów, zwracanych wartości oraz opisu typów przekazywanych danych. Do wysyłania komunikatów można użyć dowolnego protokołu wyższej warstwy, jednak zaleca się stosowanie popularnego HTTP. Sposób działania SOAP jest prosty: klient wywołuje metodę obsługiwaną przez zdalny serwer za pomocą odpowiednich instrukcji XML, po czym otrzymuje odpowiedź zapisaną w języku XML.

Struktura komunikatu SOAP [13] jest następująca:

1. Opakowanie (*envelope*) – każdy komunikat powinien zawierać jeden węzeł tego typu. Odpowiada on za opis zawartości komunikatu i sposób jego przetwarzania.
2. Nagłówek (*header*) – nagłówek jest elementem opcjonalnym. Pozwala na tworzenie dodatkowych atrybutów związanych z komunikowaniem, ale niezależnych od jego treści. Każdy komunikat może zawierać wiele nagłówków.
3. Treść (*body*) – treść komunikatu to odpowiednio sformatowane dane oraz nazwa wywoływanej metody. Sposób w jaki treść zostanie sformatowana, jest zależny od definicji wywołwanego obiektu oraz metod oferowanych przez wybrany serwer. Jeśli komunikat stanowi odpowiedź, to w jego treści może znaleźć się informacja o błędzie, która jest dodawana, gdy podczas wykonywania funkcji wystąpi błąd. Komunikaty mogą przenosić dowolne typy danych, możliwe jest budowanie różnego rodzaju złożonych struktur i nie ma właściwie żadnych ograniczeń co do liczby i rozmiaru przekazywanych parametrów.

SOAP tworzy wyższą warstwę protokołu TCP i do jego działania wystarczy port 80 wykorzystywany przez serwer WWW. Umożliwia to komunikację przez firewall'e. Zapory ogniowe mogą filtrować komunikaty SOAP'a, opierając się na nazwie obiektu, metody lub też uwzględnić oba te kryteria. Istotnym elementem bezpieczeństwa jest identyfikacja użytkownika, bowiem udostępniając usługi w Internecie, zezwalamy na korzystanie z nich w dowolny sposób. Aby ograniczyć dostęp do konkretnych usług, wystarczy zastosować kod (np. w postaci parametru wywołania usługi) umożliwiający dostęp tylko autoryzowanemu użytkownikowi. Bezpieczeństwo transmisji zapewnia SSL użyty w niższych warstwach protokołu. IBM zaproponował wprowadzenie SOAP Security Extensions, które dodają podpis cyfrowy XML, co pozwala na dodatkowe uwierzytelnienie i szyfrowanie wiadomości. Dzięki takim rozwiązaniom usługi sieciowe wykorzystujące SOAP są bezpieczne i łatwe do wdrożenia.

4.2.1. WSDL (WEB SERVICES DEFINITION LANGUAGE)

Dotychczas, gdy użytkownik chciał skorzystać z komponentów na odległym serwerze, konieczne było otrzymanie od jego operatora dokładnej specyfikacji funkcji oferowanych przez wybrany system. Obecnie problem ten rozwiązuje technologia WSDL. Jest to język umożliwiający twórcom dokładne opisanie funkcji oferowanych przez ich usługę oraz sposobu jej wykorzystania. WSDL jest instancją języka XML umożliwiającą tworzenie opisów usług sieciowych spełniających powyższe wymagania.

Aby zapytać wybraną usługę o funkcje, jakie udostępnia, należy użyć URL'a: <http://localhost/webservice.asmx?wsdl>. Odpowiedzią powinien być plik w formacie WSDL. Kompletny opis usługi musi zawierać następujące elementy [14]:

- nazwy poszczególnych usług sieciowych;
- definicję każdej operacji w zakresie używanych komunikatów;
- definicję słownictwa XML używanego w tych komunikatach;
- określenie co najmniej jednego protokołu transportu dla każdej operacji (np.: HTTP, FTP).

4.2.2. UDDI (UNIVERSAL DESCRIPTION, DISCOVERY AND INTEGRATION)

Po utworzeniu opisu w WSDL, można go udostępnić w sieci przy pomocy usługi UDDI oraz protokołu DISCO (*Discovery of Web Services*).

DISCO ułatwia odnajdywanie Web Services w sieci. Witryna publikuje dokument DISCO zawierający adresy URL opisów udostępnianych usług w formacie WSDL. Dokument DISCO zawiera odnośniki do innych witryn, a także innych dokumentów DISCO, co umożliwia przeszukiwanie drzewa katalogów.

Z kolei UDDI udostępnia użytkownikom mechanizm dynamicznego wyszukiwania innych usług sieciowych. Struktura UDDI ma postać baz danych, w których można rejestrować własne oraz wyszukiwać inne usługi sieciowe. UDDI jest rejestrem o zasięgu globalnym, a jego warstwa znajduje się nad protokołem SOAP, dzięki czemu komunikaty UDDI są opakowane w komunikatach SOAP. Sposób wyszukiwania usług za pomocą UDDI zbliżony jest do korzystania z wyszukiwarki internetowej.

Istnieje już kilka rozwiązań pozwalających na uruchomienie własnego węzła UDDI. Z reguły implementacje katalogów UDDI oferują dostęp zarówno przez strony WWW, jak i przez protokół SOAP – dzięki temu przyszłe systemy będą mogły automatycznie wyszukiwać potrzebne im usługi. W UDDI istotna jest kategoryzacja firm – musi ona umożliwiać wyszukanie firmy o ściśle określonym profilu działalności. Specyfikacja UDDI [8] opisuje również sposób wymiany danych pomiędzy serwerami, które mogą tworzyć sieć węzłów.

5. PORÓWNANIE TECHNOLOGII CORBA I SOAP

CORBA jest technologią zaimplementowaną praktycznie dla każdej platformy i wspierającą prawie każdy język programowania. Użytkownicy stosujący ten standard mogą się jednak natknąć na trudności wynikające z faktu, że nie wszystkie ORB'y potrafią się ze sobą komunikować. Potęgą CORBA tkwi w jej niezawodności i skalowalności – systemy stosujące tę technologię zapewniają efektywniejsze wykorzystanie komputerów i większą stabilność pracy.

SOAP jest znacznie łatwiejszy w implementacji, a czas jego wdrożenia jest znacznie krótszy od tego, jaki jest potrzebny w przypadku CORBY. Komunikaty SOAP'a, mają postać dokumentów XML. Jest to cecha, która – zależnie od sytuacji – może być tak zaletą, jak i wadą. Wynika to z faktu, że komunikaty w postaci XML są zwykle obszerne, w związku z czym ich przesyłanie trwa dość długo. Wymagają one również zapewnienia większej przepustowości łącza, niż odpowiadające im binarne komunikaty CORBY. W dodatku komunikaty SOAP po przesłaniu muszą być przekształcone w postać binarną, zrozumiałą dla aplikacji. Rozwiązaniem tego problemu mogłoby być wprowadzenie do aplikacji analizatorów składni XML. Wymagałoby to jednak dodatkowych kosztów związanych z rozbudową aplikacji oraz ze zwiększeniem obciążenia jednostki CPU.

Web Services i ich katalogi tworzą nowy kanał dystrybucji informacji, usług i towarów poprzez Internet, z którego korzystać mogą klienci i kontrahenci firm. Technologia ta ma jeszcze przed sobą długą drogę rozwoju. Coraz częściej pojawiają się opinie, że samo zastosowanie WSDL i UDDI nie stworzy rynków elektronicznych, a jedynie umożliwi dostęp do prostych usług. W standardach WSDL i UDDI brakuje opisu procesów biznesowych, nie można też określić kolejności wykonywanych operacji, np. związanych z zamawianiem produktu. Pojawiają się nowe propozycje standardów mających wypełnić tę lukę, np. WSFL (*Web Services Flow Language*) firmy IBM.

W tabeli poniżej (tabela 5.1) zestawione zostały najważniejsze cechy obu technologii.

Tabela 5.1

Porównanie technologii CORBA i SOAP

	<i>CORBA</i>	<i>SOAP</i>
język programowania	dowolny	dowolny
protokół komunikacyjny	IIOP	http
integracja z bazą danych	ręczna	ręczna
dostęp asynchroniczny (kolejkowy)	nie	tak
adresowanie	własna usługa <i>Naming Service</i> (niezbędne jest jej ciągłe działanie)	poprzez URI
	wskaźniki	adresowanie wykorzystujące podstawowe usługi Internetu
postać wiadomości	format binarny	format tekstowy, zbudowany na XML
		elastyczny
	zmiana w definicji interfejsu obiektu (IDL) może pociągać za sobą potrzebę rekompilacji	możliwość zamieszczania załączników
		wiadomości obszerne, długo przetwarzane
implementacja	Trudna	łatwa
czas wdrożenia	Długi	krótki

Skuteczność każdego z opisanych powyżej rozwiązań zależy w znacznej mierze od jego zastosowania. CORBA spisuje się lepiej w przypadkach, gdy ważna jest prędkość usług internetowych (np. obciążone systemy przemysłowe). Z kolei SOAP, wykorzystujący język XML, dostarcza łatwego w obsłudze interfejsu – CORBA wymaga tutaj znajomości języka IDL. Ponadto zastosowanie w SOAP akceptowanego przez firewall'e protokołu HTTP umożliwia bezproblemową współpracę z tymi powszechnie stosowanymi systemami zabezpieczeń. Jedyny problem może tu wynikać ze zbyt dużego obciążenia portu 80 (wykorzystywanego przez HTTP).

Dobrym pomysłem byłoby połączenie możliwości obu standardów: niezawodności i skalowalności CORBY z łatwością implementacji i wdrożenia SOAP'a. Przystosowanie WS do współpracy z obiektami CORBA wymaga jednak rozwiązania problemu jak odwzorowywać wiadomości SOAP na zapytania IIOP.

6. PODSUMOWANIE

Skala prowadzonych prac pozwala przypuszczać, że sieć semantyczna oraz związane z nią usługi nie pozostaną jedynie w sferze akademickich rozważań. Istnieje duża szansa, że w niedalekiej przyszłości WS będą powszechnie stosowane w życiu codziennym. Na razie jednak człowiek jest niezbędnym ogniwem takiego systemu – bez jego udziału nie może odbyć się integrowanie usług, czy praca programistyczna. Inne niedogodności to np. brak gwarancji ciągłości dostępu do usług, niezgodność opracowanych dotychczas rozwiązań, szczególnie w początkowej fazie ich rozwoju, a także brak opracowanej pro-

cedury działań w momencie wystąpienia błędów. Wreszcie wspomniany już brak zabezpieczeń oraz powszechnie uznanych standardów – specyfikacje technologii związanych z WS są cały czas zmieniane – stanowi poważną barierę hamującą rozwój tego rodzaju usług.

Warto zauważyć, że badania nad siecią semantyczna, a właściwie ich szybki rozwój, może pozytywnie wpłynąć na stworzenie standardów dla usług sieciowych. Z drugiej strony postęp w tworzeniu WS wspiera prace nad przyszłą formą sieci WWW. Dlatego tak ważną kwestią jest stworzenie uniwersalnych standardów w obu tych dziedzinach.

BIBLIOGRAFIA

- [1] Berners-Lee T.: *Semantic Web Road Map*, W3C Design Issues, October 1998.
- [2] Resource Description Framework Model and Syntax Specification, W3C Recommendation (www.w3c.org), 22 February, 1999.
- [3] Staab S.: *Emergent Semantics*, IEEE Intelligent Systems, pp. 78-86, 2002.
- [4] Fensel D.: *Ontologies: Silver Bullet for Knowledge Management and E-commerce*, Springer-Verlag, 2000.
- [5] Ushold M.: The Enterprise Ontology, The Knowledge Engineering Review, pp. 31-89, Cambridge University Press, 1998.
- [6] *Web Services Architecture Requirements*, W3C Working Draft, (<http://www.w3.org/TR/2002/WD-wsa-reqs-20021114>), 14 November, 2002.
- [7] *Web Services Description Requirements*, W3C Working Draft, (<http://www.w3.org/TR/2002/WD-wsa-desc-reqs-20021028>), 28 October, 2002.
- [8] *Oasis UDDI Specification*, (<http://www.oasis-open.org/committees/>).
- [9] *CORBA Web Services. Initial Joint Submission*, (<ftp://ftp.omg.org/pub/docs/orbos/01-06-07.pdf>).
- [10] Gokhale A., Kumar B., Sahuguet A.: *Reinventing the Wheel? CORBA Web Services*, The Eleventh International World Wide Web Conference, Hawaii, USA, 7-11 May, 2002.
- [11] Mowbray T., Malveau R.: *CORBA Design Patterns*, John Wiley & Sons, Inc., 1997.
- [12] Wallnau K.: *Common Object Request Broker Architecture. Technology Review*, Carnegie Mellon Software Engineering Institute, (<http://www.sei.cmu.edu/str/description/corba.html>)
- [13] *Web Services. Technical Overview*, Sun Microsystems, (http://dcb.sun.com/practices/webservices/overviews/overviews_soap.jsp)
- [14] *Web Services Description Language (WSDL) 1.1*, W3C Note, (<http://www.w3c.org/TR/wsdl>), 15 March, 2001.

SERVICES MANAGEMENT IN A FUTURE WEB SYSTEMS

Summary

The paper introduces the idea of a future World Wide Web which enhances content with formal semantics and machine-understandable metadata. The problems of building and managing Web Services and methods of defining Web Services are presented. The ways to establish communication between applications and informatic systems working on different platforms are specified. The paper analyses the possible specifications of the most popular technologies associated with managing the Web Services (SOAP, CORBA). Conclusions on effectiveness of presented Web Services discovering mechanisms are addressed.

Marek Moszyński, Jerzy Demkowicz, Andrzej Partyka

Katedra Systemów Geoinformatycznych, Politechnika Gdańska

INTERNETOWY SYSTEM CZASU RZECZYWISTEGO DO AKWIZYCJI I WIZUALIZACJI OBRAZÓW RADAROWYCH

Streszczenie

Obserwowana w ostatnich latach łatwość dostępu do Internetu skłania do udostępniania w sieci nie tylko szeroko rozumianych dokumentów elektronicznych, ale także wysoko specjalizowanych urządzeń. W tym przypadku pod pojęciem udostępnianie rozumie się najczęściej umożliwienie sterowania urządzeniem, jak i wizualizację danych pochodzących z tego urządzenia. W artykule tym przedstawiono realizację zdalnego interfejsu użytkownika do radaru firmy SITEX Marine Electronics Inc. RadarPC jest urządzeniem autonomicznym i może współpracować z dowolnym urządzeniem zewnętrznym poprzez zaprojektowany w tym celu protokół komunikacyjny. Dwukierunkową komunikację z radarem (sterowanie, odbiór obrazów radarowych) umożliwiają asynchroniczne porty szeregowo USB, RS422 lub RS232. Na potrzeby udostępniania pomiarów wykonywanych przez radar zrealizowano dedykowany serwer Internetowy, który realizuje na bieżąco przesyłanie skompresowanych obrazów radarowych do klientów przez sieć. W ten sposób została zrealizowana wizualizacja obrazów radarowych w aplecie pracującym w standardowej przeglądarce WWW, a ponadto dla uprzywilejowanego użytkownika udostępniono interfejs do jego sterowania.

1. WSTĘP

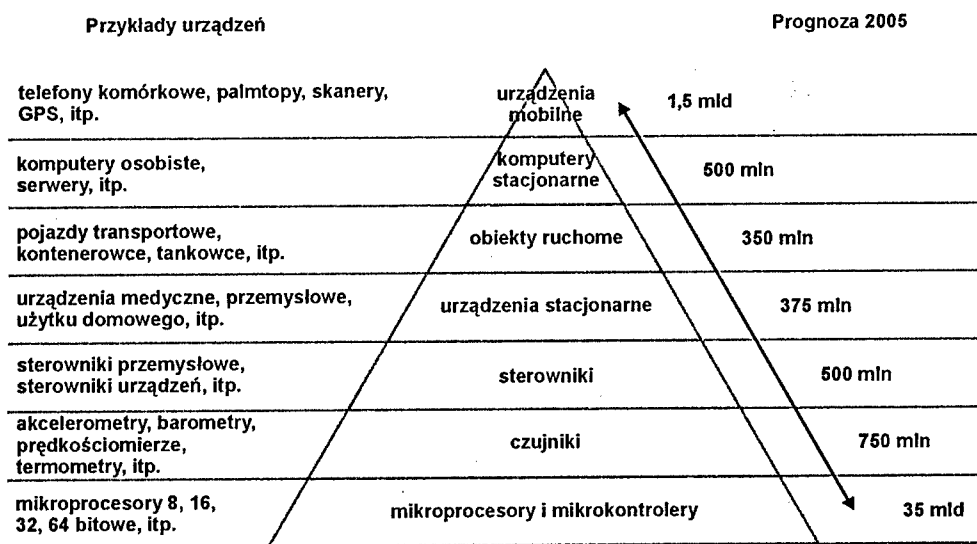
Przełom technologiczny polegający na zmianie sposobu dyfuzji informacji ma decydujący wpływ na wzrost jej podaży i popytu. Jest to główny czynnik prowadzący do informatyzacji społeczeństw, a w konsekwencji powstania społeczeństwa informacyjnego, w którym głównym przedmiotem wymiany jest informacja. Za główne medium komunikacyjne, które w przeważającej części przyczynia się do zaistnienia tej sytuacji uważa się Internet. U podłoża tych przemian leżą rewolucyjne zmiany w dziedzinie informatyki i telekomunikacji.

Według badań *Harbor Report Inc.* [1] głównymi użytkownikami Internetu w przyszłości będą urządzenia wbudowane, wymieniające informacje w różnych obszarach i dziedzinach wiedzy. Dziedziny, w których te urządzenia będą występować, można podzielić na kilka obszarów, min.: budownictwo, elektronika użytkowa, przemysł, medycyna, energetyka, transport, handel, administracja, bezpieczeństwo itp. Urządzenia wbudowane, są wysoko specjalizowanymi modułami (najczęściej mikroprocesorowymi) o dedykowanej funkcjonalności i ograniczonej ingerencji z zewnątrz.

W wymianie informacji uczestniczyć będą również urządzenia specjalistyczne, cechujące szeroko rozumiany monitoring i telemonitoring. Szczególnym rodzajem urządzeń występujących w tej dziedzinie są urządzenia pracujące w czasie rzeczywistym lub prawie rzeczywistym (*hard, soft real time*). Oczywiście w zależności od charakteru informacji, jej wagi z punktu widzenia użytkowego, należy zapewnić mechanizmy bezpieczeństwa i niezawodności jej przesyłu (szyfrowane łącze, dublowane media, łącza itp.).

Bardzo specyficznym w tym kontekście wydaje się urządzenie radarowe, RADAR (*Radio Detection and Ranging*), którego zasada działania i bardzo rozbudowany fundament teoretyczny ma duży wpływ na rozwój nie tylko metod radio detekcji, ale również innych dziedzin zdalnego monitoringu.

Najpowszechniej stosowanym radarem jest monostatyczny radar nawigacyjny, o ustalonym reżimie pracy – nadawanie/odbiór. Niestety, radar nie jest zasobem o współdzielonym charakterze. Jest zasobem o cechach urządzenia czasu rzeczywistego, któremu jednak nie można przerywać pracy. Zasobem charakteryzującym się kolejkowym sekwencyjnym dostępem, przeciwnie do Internetu. Asynchroniczność pracy radaru podobną do asynchroniczności w dostępie do informacji i jej współdzielenia w Internecie, można uzyskać jedynie poprzez wprowadzenie interfejsu (bufora), który będzie odpowiadał za translację komunikacji asynchronicznej na synchroniczną. Jest to kwestia, która decyduje o charakterze jego udostępniania w Internecie.

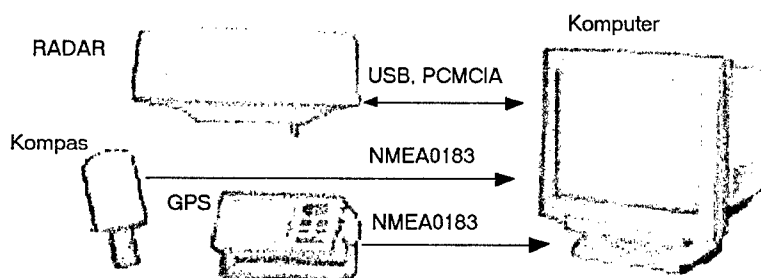


Rys.1. Wykorzystanie Internetu według prognoz Harbor Research Inc. [1]

2. SPECYFIKACJA URZĄDZENIA RADAR-PC

W eksperymencie wykorzystano Radar-PC firmy SI-TEX, który jest przykładem nowego trendu w tej dziedzinie. Do tej pory radar był urządzeniem centralnym, ośrodkującym prace innych urządzeń peryferyjnych. Radar integrował wszystkie urządzenia i przetwarzał otrzymaną od nich informację. Aktualnie, ciężar obliczeń przejmie komputer stacjonarny

– jednostka centralna. Do komputera mogą być podłączone dowolne źródła informacji, w tym radar, poprawiając lub wspomagając tym samym proces poprawnej interpretacji i czytelność akwizowanej informacji radarowej (rys. 2).

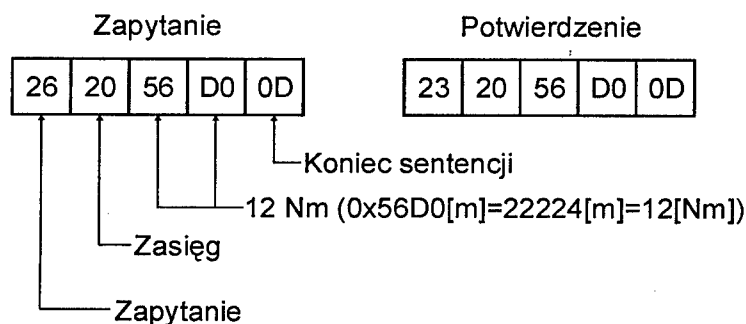


Rys. 2. Schemat funkcjonalny systemu Radar-PC

Radar-PC składa się z dwóch zasadniczych części: skanera i jednostki wyświetlającej obrazowanie radarowe (komputera z odpowiednią aplikacją, komunikującą się z radarem) rys. 2. Skaner radarowy składa się z anteny nadajnika i odbiornika (*tranceiver*) oraz jednostki komunikującej się z komputerem.

Najważniejszymi parametrami radaru są jego moc i zasięg (w tym wypadku 2 kW i 16 Nm), częstotliwość pracy (9445+/-30 MHz) oraz minimalny/maksymalny czas trwania impulsu (0.1 μ s/2200 Hz i 0.8 μ s / 50 Hz). Komunikacja radaru z komputerem realizowana jest z wykorzystaniem specjalnego protokołu, poprzez złącze RS-422, dołączane przez translatory do portu USB. Unikalny protokół, zbliżony jest do formatu SeaTalk[®] firmy Autohelm\Raytheon (rys. 3). Każda sentencja sterująca tego protokołu wymaga potwierdzenia. Pierwszy bajt w sentencji oznacza pytanie lub odpowiedź, następny oznacza rodzaj komunikatu sterującego, kolejne dwa są argumentami komunikatu, ostatni bajt jest znacznikiem końca sentencji.

Protokół umożliwia zdalne sterowanie radarem oraz odbiór dynamicznie zmieniającego się obrazu radarowego w postaci cyfrowej co stanowi unikalną cechę radaru cyfrowego.



Rys. 3. Przykład sentencji sterującej zmianą zakresu radaru

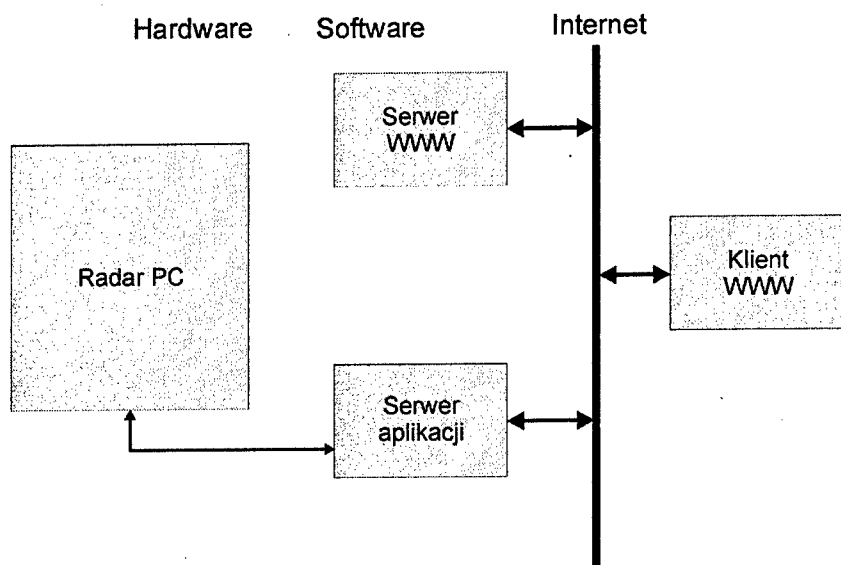
RS-422 jest standardem przemysłowym umożliwiającym przesyłanie danych w sposób szeregowy i asynchroniczny (możliwa jest transmisja dwukierunkowa i jednokierunkowa, simplex, duplex oraz tzw. *multi point*) na dużych odległościach, z maksymalną szybkością transmisji (12m/10Mb/s – 1200m/100Kb/s). W przypadku Radar-PC z prędkością 115200 bps. Różnicowy sposób przesyłania danych zapewnia dużą odporność na zakłócenia zewnętrzne.

Do podstawowych funkcji sterujących systemem RadarPC należą: możliwość zmiany zakresu, czułości radaru (*Gain*), zapamiętywanie ustawień dla danego zakresu pracy (*Keep Range Setting*) oraz funkcja *playback* (*Real Time Recording*). W zakres funkcjonalności nawigacyjnych wchodzi min.: elektroniczna linia namiaru (EBL), elektroniczny znacznik odległości (VRM), wprowadzanie stref niebezpiecznych (*guard zones*), echo radarowe (*Radar Trials*). Funkcje poprawiające efekt zobrazowania w różnych warunkach meteorologicznych to min.: STC (*Sea Clutter Control*), FTC (*Fast Time Constant*).

Obraz radarowy jest zintegrowany z mapą elektroniczną. Obydwie technologie uzupełniają się w procesie interpretacji zdalnie akwizowanego obrazu radarowego.

3. ARCHITEKTURA SYSTEMU

Udostępnianie obrazów radarowych w sieci Internet zrealizowano według klasycznej już dziś architektury klient-serwer przedstawionej na rys. 4. Podobną architekturę autorzy wykorzystali już w poprzednio zrealizowanym projekcie udostępniającym miniaturową echosondę cyfrową w sieci Internet [2]. W trakcie implementacji zrealizowano kilka wersji zależnych od docelowej platformy sprzętowej i programowej. Dotyczy to głównie oprogramowania pracującego po stronie serwera.



Rys. 4. Architektura klient-serwer zastosowana do udostępniania obrazów radarowych w sieci Internet

Pierwsza wersja oprogramowania po stronie serwera została wykonana jako wielowątkowa i została zorganizowana w jednym module programowym implementującym serwer WWW i serwer aplikacji. W omawianym rozwiązaniu serwer WWW jest serwerem iteracyjnym obsługującym jednego klienta w danym momencie. Taka minimalna implementacja związana jest z jedyną funkcją jaką pełni serwer WWW; służy on bowiem wyłącznie do dystrybuowania dokumentów HTML i towarzyszących im apletów, które są plikami o niewielkich rozmiarach. Serwer aplikacji natomiast jest serwerem wielobieźnym obsługującym kilku użytkowników równolegle. Jego głównym zadaniem jest implementacja protokołu komunikacyjnego z Radarem-PC, gromadzenie na bieżąco kolejnych obrazów radarowych i wysyłanie ich do zarejestrowanych klientów. Klienci ci są rejestrowani poprzez ściągnięty razem z dokumentem HTML aplet napisany w Javie. Aplet ten jednocześnie realizuje wyświetlanie obrazów radarowych w przeglądarce klienta. Oprogramowanie w tej wersji bez żadnych modyfikacji może również współpracować z regularnym serwerem WWW.

Jak się okazało, komunikacja z radarem i obsługa klientów jest zadaniem mocno obciążającym serwer. Tak więc w kolejnej wersji oprogramowania zdecydowano się na wyodrębnienie fragmentu kodu realizującego konwersję danych pobieranych z łącza RS (lub USB) na protokół zgodny z TCP. Ta koncepcja pozwoli na bezpośrednie przeniesienie fragmentu oprogramowania do urządzenia wbudowanego i zagwarantuje łatwość wykorzystania systemu w połączeniu z regularnym serwerem WWW Katedry Systemów Geoinformatycznych. Otworzy to jednocześnie możliwość przechowywania obrazów radarowych w bazie danych serwera, jako danych o charakterze historycznym.

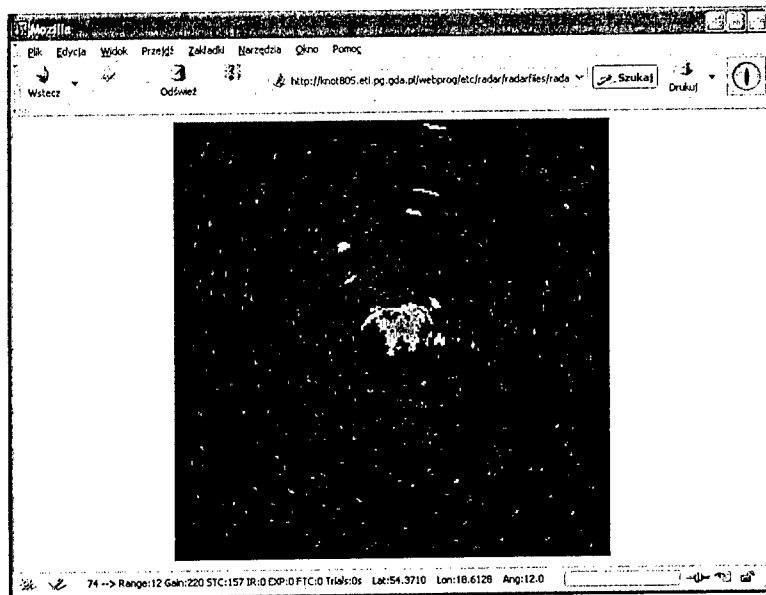
Odrębnym problemem implementacyjnym jest umożliwienie sterowania urządzeniem. Aby zapewnić "świadomy" dostęp do urządzenia umożliwiono użytkownikowi podanie hasła, weryfikowanego przez serwer aplikacji. W ten sposób możliwa jest swego rodzaju wirtualizacja radaru, gdzie każdy uprawniony użytkownik może sterować pracą urządzenia a efekt zmian obserwowany jest przez każdego obserwatora.

Oprogramowanie po stronie klienta wykorzystuje jedynie zasoby udostępniane przez przeglądarkę WWW. Dokumenty HTML dostarczane klientowi wykorzystują zarówno mechanizmy języka skryptowego JavaScript jak i komunikację pomiędzy apilem dostarczoną klientowi w postaci klasy napisanej w Javie a serwerem aplikacji. Dla przeglądarek bez Javy i Javascriptu przygotowano również dokumenty HTML z pseudo-obrazami w postaci tabeli odświeżane z częstotnością pracy radaru tj. ok. 2 sekund.

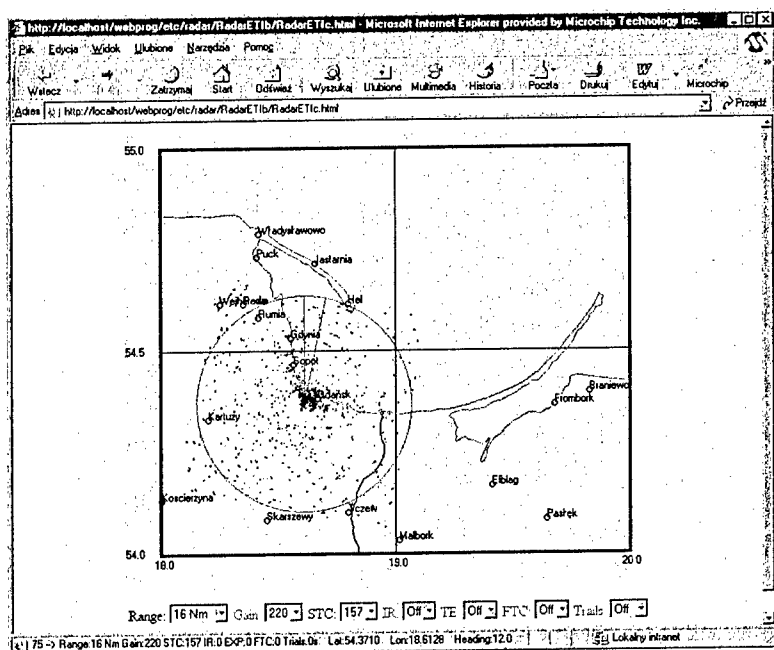
4. WYNIKI

Na rysunkach 5 i 6 przedstawiono zrzuty ekranów ilustrujące funkcjonowanie systemu. Rysunek 5 pokazuje obraz radarowy wyświetlany przez aplet zanurzony w dokumencie HTML. Oryginalny obraz radarowy jest czterokolorowym obrazem rastrowym o wymiarze 240x240 pixeli zajmującym ok. 14kB wraz z nagłówkiem. Przesyłany on jest w postaci skompresowanej i w zależności od zawartości zajmuje od 2-4kB.

Rysunek 6 pokazuje obraz umieszczony na warstwie uproszczonej mapy wektorowej. Dodatkowo w dolnej części umieszczono prosty interfejs umożliwiający sterowanie radarem.



Rys. 5. Przykład obrazu radarowego pochodzącego z radaru umieszczonego na dachu budynku Wydziału Elektroniki i Telekomunikacji Politechniki Gdańskiej prezentowanego w przeglądarce Mozilla PL.



Rys. 6. Obraz radarowy zlokalizowany geograficznie na tle prostej mapy wektorowej wraz z interfejsem dostępnym dla uprawnionego użytkownika do sterowania parametrami radaru prezentowany w przeglądarce Microsoft Internet Explorer.

5. WNIOSKI

Prezentowany system jest propozycją nowej aplikacji tradycyjnego urządzenia radarowego. Dzięki temu obrazy radarowe mogą być udostępnione szerokiemu kręgowi użytkowników Internetu.

System pozwala na przeprowadzenie badań związanych z możliwościami technologicznymi udostępniania urządzeń radarowych w Internecie oraz zbadania zapotrzebowania na tego typu usługi.

Badania dotyczące BHP urządzenia radarowego zostały przeprowadzone przez Urząd Radiokomunikacji.

BIBLIOGRAFIA

- [1] *The Pervasive Internet Opportunity*, Harbor Research Inc., 2003
- [2] Moszyński M., Stepnowski A.: *Przenośna echosonda cyfrowa z prezentacją w sieci internet*, Technologie Informacyjne, Gdańsk, 2003

REAL TIME SYSTEM FOR RADAR ACCESS IN THE INTERNET

Summary

The development of multi-aspect systems running as an Internet applications open the gates also for distribution of specialized devices in the Internet. In this case, the distribution of device means its virtualization among many clients spread in the net. The paper presents the concept of virtualization of commercially available radar RadarPC produced by SITEX Marine Electronics Inc. RadarPC is autonomous device, which could communicate with other computer devices using specialized software protocol running over serial standard protocols like RS232, RS422 or USB. The distribution of data produced by the radar has been achieved by development of dedicated WWW server, that provides real-time compression of data acquired by the radar and its distribution to logged clients. The client applet downloaded from the server allows for visualization of radar images in the classical WWW browser, and it enables control of the device for authorized users.

Bartosz Paliświat, Jerzy Nawrocki

Instytut Informatyki, Politechnika Poznańska

WARSTWA PREZENTACJI W WIELOWARSTWOWYCH SYSTEMACH INFORMACYJNYCH*

Streszczenie

Właściwie każdy współczesny system informacyjny, niezależnie od jego wielkości, tworzony jest w oparciu o architekturę wielowarstwową. O ile rozwiązania z wewnętrznych warstw logiki biznesowej czy też np. z warstw źródeł danych są dobrze opracowane i były wielokrotnie przedstawiane w różnych artykułach, o tyle warstwa prezentacji wydaje się być nieco pomijana. Dotyczy to w szczególności problemu ponownego użycia fragmentów kodu tej warstwy oraz technik jej modularyzacji.

W poniższym artykule przedstawiono i porównano najważniejsze rozwiązania pozwalające na wydzielenie w architekturze systemu warstwy prezentacji. W szczególności porównano ze sobą systemy oparte o transformację XSLT danych zapisanych przez warstwę logiki w postaci XML, typowe rozwiązania z wykorzystaniem wzorców warstwy prezentacji (z uwzględnieniem komunikacji typu push i pull) oraz różne technologie typu Server Pages. Zwrócono też uwagę na różnorodne możliwości łączenia opisanych technologii i płynące z tego korzyści oraz zagrożenia.

Jako kryteria porównania wzięto pod uwagę wspomniane już problemy związane z modularyzacją i ponownym użyciem kodu. Omówiono również ewentualne różnice w sposobie organizacji pracy zespołowej, które mogą pojawić się podczas używania każdej z tych technologii ze względu na wymienione wcześniej czynniki, a także czytelność tworzonego kodu oraz jakość rozdzielania warstwy prezentacji od logiki. Zwrócono również uwagę na problemy związane z pielęgnacją kodu tej warstwy.

1. WSTĘP

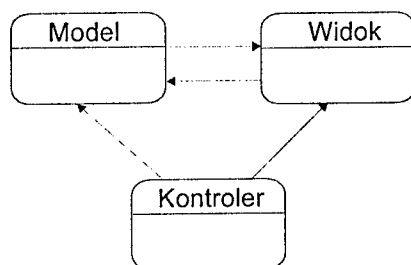
Podejścia do budowy systemów informacyjnych ewoluowały (i ewoluują) na wiele różnych sposobów. Stąd rozwiązania spotykane w dniu dzisiejszym wydają się znacząco od siebie różnić, czy to na skutek tego, iż przebyły zupełnie różne drogi, czy też dlatego, że u ich podłoża leżała motywacja zaadresowania zupełnie różnych problemów. Jednak podczas tej ewolucji wykształciły się pewne mechanizmy, które ze względu na swoją prostotę, użyteczność i szeroko rozumianą „elegancję” odcisnęły swoje piętno na niemal wszystkich

* Praca finansowana z grantu BW-91-399/04

stosowanych obecnie rozwiązaniach. O rozwiązaniach takich możemy myśleć jak o swojego rodzaju wzorcach architektonicznych, będących uogólnieniem wiedzy ludzi pracujących latami na systemami tego typu.

Dominującą architekturą wśród współczesnych aplikacji jest architektura wielowarstwowa. Podział na warstwy pojawia się na różnych poziomach architektury systemów, zwykle jednak przejawia się on rozdzieleniem warstw logiki biznesowej, warstw prezentacji i warstw związanych z przechowywaniem danych. Nie oznacza to oczywiście, że podział ten nie może pójść dalej, w sposób specyficzny dla danej aplikacji dzieląc np. warstwę logiki biznesowej na dalsze podwarstwy. Podział taki może mieć u podłoża z natury rozproszony charakter aplikacji (np. dla aplikacji opartych o architektury komponentowe), choć nie jest to oczywiście jedyna możliwość.

Głęboki związek z architekturą wielowarstwową ma paradygmat MVC (Model – View – Controller) [8]. Polega on na podzieleniu aplikacji (lub jej części) na trzy niezależne warstwy: modelu, widoku i kontrolera. To co nazywamy modelem, to nic innego jak część aplikacji odpowiedzialna za modelowanie (najczęściej w sposób obiektowy lub komponentowy) świata właściwego dziedziny danej aplikacji. Model nie zawiera kodu związanego z graficzną prezentacją wyników swojego działania. Nie zawiera też bezpośrednio kodu związanego z reakcją na wejście dostarczane przez użytkownika. Te umieszczone są odpowiednio w warstwie widoku i warstwie prezentacji.



Rys. 1. Model MVC

Oczywistą konsekwencją stosowania paradygmatu MVC jest wydzielenie z systemu warstwy prezentacji. W poniższym artykule przyjrzymy się bliżej sposobom jej separacji w aplikacjach internetowych, kładąc nacisk na sposób komunikacji pomiędzy warstwą prezentacji a warstwą (warstwami) modelu.

2. TECHNOLOGIE TYPU SERVER PAGES

Technologie typu server pages zakładają osadzenie wykonywalnego kodu wewnątrz kodu warstwy prezentacji. Kod taki wyróżniony jest zwykle przez otoczenie go specjalnymi znacznikami, pozwalającymi parserowi rozróżnić (nic dla niego nie znaczący) kod HTML, od kodu danego języka.

Założenie to wydaje się od samego początku burzyć omówione wyżej idee modelu MVC. Nie jest to jednak do końca prawda. Podejścia tego typu umożliwiają (często nawet w prosty i elegancki sposób) rozdzielenie kodu logiki od prezentacji. Niestety fakt, że to umożliwiają, nie oznacza jednak w żaden sposób, że wspierają ten styl pisania. Stąd niestety duża liczba systemów napisanych w tych technologiach zupełnie nie korzysta z tej

możliwości, co sprawia, że z czasem stają się niesłychanie trudne w utrzymaniu. Kolejną wadą jest fakt, że do napisania warstwy prezentacji wymagana jest mniejsza lub większa znajomość danego języka programowania. Znacząco utrudnia to współpracę między zespołami zajmującymi się poszczególnymi aspektami systemu (logiką i prezentacją), w szczególności wymaga od grafików chociażby znajomości podstaw danej technologii.

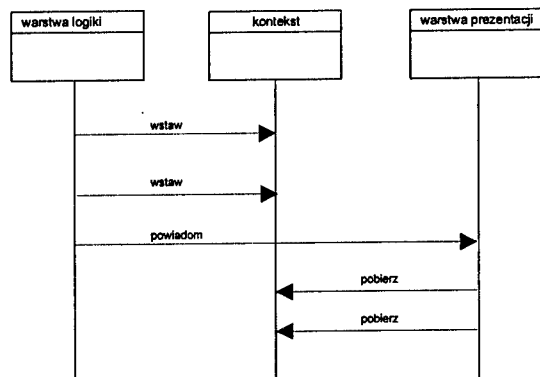
Mimo to, technologie typu server pages nie tylko są bardzo popularne w chwili obecnej, ale ich popularność bardzo szybko rośnie. Wynika to przede wszystkim z wygody, jaką oferują korzystającym z nich programistom. Nie tylko pozwalają ukryć szczegóły implementacyjne nie związane bezpośrednio z logiką pisanej aplikacji, ale także wymuszają pewien, wygodny dla wielu osób, styl myślenia, w którym przetwarzanie przebiega dokładnie tak, jak wypisywana jest strona na ekranie przeglądarki użytkownika. Ponadto języki typu server pages oferują szeroką gamę udogodnień, które pozwalają szybko, zwykle za pomocą pojedynczego wywołania instrukcji, rozwiązać najczęściej spotykane drobne problemy.

3. WZORCE WARSTWY PREZENTACJI

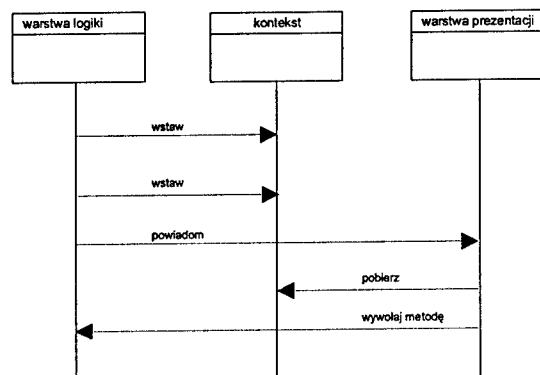
Chyba najbardziej popularnym podejściem jest jednak przesunięcie kodu HTML implementującego wygląd strony do tzw. wzorców warstwy prezentacji. Są to osobne pliki, zawierające wyłącznie kod HTML oraz kawałki prostego kodu takiego jak pętle czy odwołania do dynamicznie generowanych wartości. Warstwa logiki umieszczona jest zupełnie oddzielnie, jako np. servlet.

Kluczowym pojęciem przy tego typu podejściu jest tzw. kontekst. Jest to swoisty rodzaj pośrednika pomiędzy warstwami, do którego warstwa logiki wstawia obiekty, dostępne później za pomocą wspomnianych języków skryptowych w warstwie prezentacji. Jeżeli wszystkie dane, do których ma mieć dostęp wzorzec muszą być explicite wstawione (i w związku z tym obliczone) do kontekstu przed odwołaniem do wzorca, mówimy, że mamy do czynienia z tzw. podejściem typu „push”. Podejście odmienne, polegające na tym, że z poziomu języka skryptowego programista ma możliwość wywołania metody na obiekcie wstawionym do kontekstu w celu pozyskania potrzebnych danych nazywa się podejściem typu „pull”. Należy zauważyć, że podejście typu „pull” występuje zazwyczaj łącznie z podejściem typu „push”. Idee tych podejść zaprezentowano na diagramach sekwencji na rysunkach 2 i 3.

Zastosowanie wzorców warstwy prezentacji ma wiele zalet. Najważniejszą z nich jest tak ściśle jak tylko możliwe rozdzielenie pomiędzy warstwą prezentacji a logiką aplikacji. Wcześniejsze zdefiniowanie zmiennych potrzebnych podczas wyświetlania rezultatów działania systemu powoduje, że prace zespołów programistów i grafików mogą być prowadzone niemal zupełnie niezależnie. Przyczynia się do tego przede wszystkim prostota stosowanych we wzorcach języków skryptowych, która powoduje, że nawet osoby nie związane bezpośrednio z programowaniem mogą bez trudu przyswoić podstawy ich składni (choć zwykle nawet to nie jest konieczne).



Rys. 2. Podejście typu "push"



Rys. 3. Podejście typu łączonego ("push" i "pull")

Tak naprawdę o możliwościach podejścia z warstwami prezentacji decydują jednak w znacznym stopniu możliwości stosowanych do przetwarzania wzorców narzędzi, a co za tym idzie możliwości stosowanych języków skryptowych. To od tych języków zależy jak czytelne będą wzorce, jak łatwo będzie je tworzyć i czy będą się one nadawać do ponownego użycia. Dlatego w części poświęconej narzędziom porównanie silników umożliwiających wykorzystanie podejścia opartego o wzorce warstwy prezentacji opierać się będzie głównie na porównaniu dostarczanych przez nie języków.

Do rozwiązań opartych na wzorcach warstwy prezentacji możemy zaliczyć także systemy, w których komunikacja pomiędzy modelem a warstwą prezentacji odbywa się za pośrednictwem języka XML. Wzorce prezentacji mają wtedy postać transformacji XSLT, która przekształca zawartość kontekstu do postaci czytelnej dla użytkownika. Niestety niewielka czytelność takiego rozwiązania w połączeniu z ograniczonymi możliwościami sprawiają, że rozwiązanie to zdecydowanie ustępuje popularnością opisanym w rozdziale 5.

4. NARZĘDZIA DLA TECHNOLOGII SERVER PAGES

Ze względu na szerokie spektrum rozwiązań, często znacząco od siebie różnych, nie sposób mówić o możliwościach technologii server pages bez powoływania się na konkretne języki. Poniżej omówiono dwa najbardziej popularne i bardzo dynamicznie rozwijające się: JSP [4] i PHP [2,6].

Język JSP powstał na bazie technologii Java Servlets. Jest on tak naprawdę po prostu innym sposobem zapisu servletów, gdyż skrypty JSP tłumaczone są przed kompilacją właśnie na servlety Javy. Z logicznego punktu widzenia, zapis ten jest jednak zupełnie różny. O ile (abstrahując od opisywanego niżej mechanizmu wzorców) pisanie servletów polega na osadzaniu kodu HTML wewnątrz kodu Javy, o tyle pisanie w JSP oznacza osadzanie kodu Javy wewnątrz stron HTML. Zapewne nie byłoby w tym nic szczególnego, gdyby nie fakt, że JSP bardzo ściśle współpracuje ze standardem Java Beans. Podejście to, raczej nietypowe dla technologii server pages, pozwala na oddzielenie warstwy logiki od kodu warstwy prezentacji. Co więcej, narzucenie używanym klasom Javy zgodności ze standardem Java Beans nie tylko poprawia czytelność kodu, ale również ułatwia ewentualne rozproszenie aplikacji i stanowi doskonałe wsparcie dla ponownego użycia takich klas. Drugim mechanizmem przyczyniającym się do wzrostu popularności JSP jest możliwość definiowania własnych znaczników i grupowania ich w biblioteki. Zastosowanie takiego mechanizmu przyczyniło się do powstania przeróżnych bibliotek takich znaczników co w oczywisty sposób uprościło pisanie aplikacji za pomocą JSP.

PHP jest znacząco różne od JSP. Przede wszystkim korzysta z własnego, specyficznego języka, dużo uboższego od Javy, wyposażonego jednak w pewne specyficzne dla tworzenia stron internetowych mechanizmy, znacząco przyspieszające pisanie aplikacji w tym języku. O sile PHP stanowi jednak nie język, ale bogactwo zewnętrznych modułów, które rozszerzają zasób dostępnych w PHP funkcji. W odróżnieniu od funkcji definiowanych przez użytkownika, są one dostarczane w postaci skompilowanej (dla konkretnej platformy sprzętowo – programowej), dzięki czemu są one wykonywane znacznie szybciej. Obecnie rozwój PHP idzie w kierunku rozbudowy ubogich obecnie mechanizmów obiektowych, a także integracji z innymi technologiami (dostępne są moduły umożliwiające wykorzystanie klas Javy czy np. korzystanie z Web Services). Mimo, iż nie wydaje się konieczne stosowanie osobnych narzędzi do wspierania podejścia opartego na wzorcach warstwy prezentacji, narzędzia takie są dostępne. Ideą ich powstania było dyscyplinowanie programistów, którzy nadużywając swobody jaką dają technologie server pages, często swobodnie mieszały ze sobą prezentację i logikę aplikacji. Niestety narzędzia te nie są ani zbyt wydajne, ani popularne. Powoduje to, że mimo iż można napisać w PHP kod, który będzie czytelny i zupełnie zgodny z modelem MVC, zwykle użycie tego języka w większych projektach prowadzi (prędzej czy później) do bałaganu w kodzie i pełnego przemieszania warstw.

5. NARZĘDZIA DLA WZORCÓW WARSTWY PREZENTACJI

W rozdziale tym porównano dwa narzędzia pozwalające na stosowanie podejścia opartego na wzorcach warstwy prezentacji w aplikacjach opartych o Java Servlets: Velocity [5] i Freemarker [7]. Oba narzędzia prezentują podobną funkcjonalność, występują w nich jednak pewne warte podkreślenia różnice.

Istotną cechą Velocity jest to, że umożliwia realizację podejścia typu push + pull. Velocity udostępnia język skryptowy (VTL – *Velocity Template Language*), z którego

poziomu można odwoływać się do obiektów umieszczonych wcześniej w kontekście. Obejmuje to nie tylko pobranie ich wartości w postaci prostego łańcucha znaków, ale również odwołanie do tzw. własności (co sprowadza się do wywołania odpowiedniej metody zgodnie ze standardem Java Beans) oraz po prostu bezpośrednie wywołania metod.

Zasadniczą zaletą języka skryptowego udostępnianego przez Velocity jest jego lekkość i prostota. Oprócz standardowych konstrukcji takich jak pętle czy instrukcje warunkowe, udostępnia on również prosty mechanizm makrodefinicji, pozwalający czy to na stworzenie lokalnie wykorzystywanych w ramach pojedynczego wzorca makr, czy też utworzenie całej biblioteki makrodefinicji w celu jej rozpowszechniania lub po prostu użycia w wielu projektach. Język ten posiada również sporo ułatwień dla programistów, które niestety czasami mogą prowadzić do niejednoznaczności i pomyłek. Samo narzędzie ma również bardzo mocne wsparcie ze strony innych projektów.

Dużo większe możliwości od Velocity oferuje (w nowych wersjach) Freemarker. Posiada on zdecydowanie bardziej rozbudowany mechanizm makrodefinicji, umożliwiający m.in. definiowanie za ich pomocą własnych znaczników. Makrodefinicje mogą posiadać zmienne lokalne, co w połączeniu z możliwością rekursywnego wywoływania daje programiście dużo ciekawych możliwości. Język (FTL – Freemarker Template Language) posiada również zaawansowane mechanizmy, służące do ewaluacji fragmentów wzorca do zmiennej oraz traktowania zmiennych jako fragmentów wzorca. Posiada również wiele ułatwień dla programistów, związanych z obsługą lokalizacji (formaty czasu, liczb zmienno-przecinkowych), a także mechanizm rejestracji błędów.

Wiele mechanizmów FTL zaprojektowano z myślą o ponownym wykorzystaniu kodu. Makrodefinicje można łączyć w biblioteki (podobnie jak w Velocity), jednak Freemarker wprowadza dodatkowo pojęcie przestrzeni nazw, co znacząco ułatwia wykorzystanie bibliotek pochodzących od różnych dostawców. Co więcej, FTL potrafi współpracować z bibliotekami znaczników JSP, a nawet operować na obiektach Pythona.

Freemarker stosuje wyłącznie podejście typu push. Nie oznacza to jednak, że nie można za jego pomocą odwoływać się do klas Java Beans. Nie jest to jednak tak proste jak w przypadku Velocity. Najczęściej stosowaną praktyką w projektach używających Freemarkera jest po prostu przekazywanie wszystkich potrzebnych wartości za pomocą wartości prostych oraz kolekcji.

6. ZAKOŃCZENIE

Porównując możliwości poszczególnych podejść i narzędzi można by dojść do wniosku, że zdecydowanie najszerze możliwości oferują podejścia oparte na wzorcach warstwy prezentacji. Wniosek taki byłby jednak zbyt pochopny. Należy pamiętać, że często prostota i szybkość jaką oferuje np. PHP równoważy straty (szczególnie przy niewielkich projektach, realizowanych przez wąskie grono osób) powodowane nieco gorszą czytelnością kodu. JSP z kolei oferuje rozsądny kompromis pomiędzy jakością separacji a wygodą dalszego utrzymania kodu, który w niektórych projektach może dać doskonałe rezultaty. Należy również pamiętać, że mimo iż podejścia typu server pages nie narzucają podziału kodu na warstwę logiki i prezentacji, to jednak pisanie w tym stylu jest jak najbardziej możliwe, choć z pewnością przy większych projektach narzucenie sztywnych reguł podziału może być trudne. Stąd wybór najlepszej technologii pozostaje i z pewnością długo będzie pozostawał problemem nierozstrzygniętym, o czym świadczy chociażby bogactwo dostępnych, konkurujących ze sobą narzędzi.

BIBLIOGRAFIA

- [1] Arciniegas F., XML- Kompendium programisty, Helion, 2002
- [2] Choi W., Kent A., Lea C., Prasad G., Ullman, C.: PHP 4 od podstaw, Helion, 2002.
- [3] Paliświat B., Komponenty typu szczelina – wypełnienie w tworzeniu witryn internetowych, Zeszyty Naukowe Wydziału ETI PG, Technologie Informacyjne nr 1, str. 70 – 74, WETI PG, Zakład Poligrafii Politechniki Gdańskiej, 2003
- [4] Java Server Pages Technology, <http://java.sun.com/products/jsp>
- [5] The Apache Jakarta Project – Velocity, <http://jakarta.apache.org/velocity>
- [6] PHP home page, <http://www.php.net/>
- [7] Freemarker, <http://freemarker.sourceforge.net>
- [8] Java *BluePrints* – Model – View – Controller, <http://java.sun.com/blueprints/patterns/MVC-detailed.html>

PRESENTATION LAYER IN MULTILAYER INFORMATION SYSTEMS

Summary

Almost every information system regardless of its size, is created on the basis of multilayer architecture. In comparison with solutions from internal layers (like business logic layers or data source layers), the presentation layer seems to lag a little behind, especially in the area of code reuse and modularization.

In the following article the most important solutions, which makes possible the separation between presentation layer and business logic, are described. Especially systems based on XSLT transformations, presentation layer templates (including “push” and “pull” communication) and different server pages technologies are compared. Also possibilities of combining these technologies and benefits and threats following from it are discussed.

As comparison criteria, problems concerning modularization, reusability and maintenance were taken into consideration. Also the differences in team work organization, which appears because of factors mentioned above, but also code readability and the degree of layers separation are shortly described.

Paweł Sachse¹, Krzysztof Juszcyszyn²

¹Instytut Matematyki, ²Instytut Sterowania i Techniki Systemów
Politechnika Wrocławska

MODEL ZAUFANIA DLA SYSTEMÓW WEBOWYCH

Streszczenie

W pracy przedstawiono zagadnienie budowy tzw. sieci zaufania w rozproszonych środowiskach obliczeniowych (jak sieć WWW). Zaproponowano probabilistyczny model zaufania dla systemów webowych, omówiono możliwości jego zastosowania oraz budowy polityk zaufania dla działających w sieci podmiotów.

1. WSTĘP

Sieć WWW staje się globalnym, dynamicznym środowiskiem przetwarzania i składowania informacji, pozbawionym globalnego autorytetu, który mógłby zagwarantować bezpieczeństwo i jakość udostępnianych danych. Dotychczasowe prace koncentrowały się głównie na zapewnieniu integralności i poufności informacji, pozostawiając otwartą kwestię oceny *jakości* zasobów udostępnianych przez dany podmiot [3]. Coraz częściej potrzebujemy mechanizmu, który pomógłby odpowiedzieć na pytanie: „Czy mogę zaufać temu użytkownikowi/serwisowi/usłudze?”. Pojęcie zaufania (*trust*) definiowane jest na różne sposoby w dziedzinach takich jak nauki społeczne, badania operacyjne oraz informatyka. Ogólnie rozumiemy je jako przekonanie nt. niezawodności bądź zgodnego z oczekiwaniami zachowania podmiotu [8].

W informatyce podstawową dla większości modeli zaufania strukturą jest skierowany, etykietowany graf (tzw. graf zaufania), w którym węzły odpowiadają *agentom* (podmiotom; aktywnym komponentom systemu – użytkownikom, programom, usługom), a krawędzie – relacjom zaufania pomiędzy nimi. Krawędzie etykietowane są wartościami (poziomami) zaufania, najczęściej są to liczby rzeczywiste z przedziału $[0,1]$, bądź elementy predefiniowanego zbioru. Agenci mogą subiektywnie przypisywać etykiety krawędziom, oceniając w ten sposób zachowanie innych podmiotów w sieci. Każdy z modeli zaufania korzysta także z tzw. *metryki zaufania* (*trust metric*), której zadaniem jest oszacowanie poziomu zaufania między dwoma węzłami grafu nie połączonymi bezpośrednio krawędzią. Przykładowe metryki omówiono w [9], podczas gdy pierwsze zastosowania w środowisku systemów webowych opisano w [6] oraz [5]. Bardzo często

poziomy zaufania (liczby z $[0,1]$) interpretuje się jako prawdopodobieństwa zgodnego z oczekiwaniami zachowania.

Z drugiej strony coraz częściej wskazuje się, że sieć WWW wykazuje cechy badanych od dawna sieci społecznych (*social networks*), zarówno pod względem topologii relacji [1] jak i zachowania użytkowników [10]. W rezultacie żądana metryka zaufania dla systemów webowych powinna uwzględniać czynniki wpływające na relacje zaufania w sieciach społecznych. Zdefiniowano je w [4], znane także pierwsze zastosowania w systemach webowych [7]. Najważniejsze własności prezentowanej w tym artykule metryki zaufania są następujące:

1. Poziom zaufania między agentami zależy od historii interakcji między nimi.
2. Agenci należą do *grup*. Przynależność do grupy może skutkować określonym poziomem zaufania bez względu na wcześniejsze działania danego agenta.
3. Źródłem zaufania może także być *reputacja* (opinia zaufanych agentów).

Dodatkowo, pożądana jest niezależność od zastosowań w konkretnej dziedzinie (obszarze tematycznym) oraz możliwość wykorzystania webowych standardów zapisu i opisu informacji (jak XML czy RDF). W kolejnych rozdz. zostanie przedstawiony probabilistyczny model zaufania spełniający powyższe założenia..

2. OBLICZENIOWY MODEL ZAUFIANIA

2.1. Poziom zaufania pomiędzy węzłami grafu zaufania

Niech $A = \{a_1, a_2, \dots, a_n\}$ oznacza zbiór n podmiotów (agentów). Rozważamy *interakcję* pomiędzy podmiotami a_i i a_j (która może być jak najogólniejszej natury: może nią być np. wymiana danych, transakcja e-commerce, usługa sieciowa itp.). Po zakończeniu interakcji a_i ocenia a_j przyporządkowując mu wartość $x^{i,j}$, gdzie $x^{i,j} \in [0,1]$. Intuicyjnie $x^{i,j} = 1$ oznacza, że a_i jest w pełni usatysfakcjonowany, natomiast $x^{i,j} = 0$ równoważne jest niezadowoleniu z dokonanej usługi (niespełnieniu oczekiwań). Po pewnym czasie a_i dysponuje $m=m(i,j)$ obserwacjami wiarygodności a_j , które będziemy oznaczać $x_1^{i,j}, x_2^{i,j}, \dots, x_m^{i,j}$. Rozważmy zaufanie, jakim a_i darzy a_j , jako zmienną losową $X^{i,j}$. Uzasadnione jest to następującą obserwacją: usatysfakcjonowanie a_i z pojedynczej transakcji z a_j z pewnością będzie zmienne, jednak "natura" procesu pozostaje podobna. ($x_1^{i,j}, x_2^{i,j}, \dots, x_m^{i,j}$) jest zatem próbą losową rozmiaru m ze zmiennej losowej $X^{i,j}$. Będziemy zakładać, że $X^{i,j}$ są niezależne. Jak wiadomo, wartość oczekiwaną i wariancję $X^{i,j}$ (odtąd oznaczane odpowiednio $E(X^{i,j})$ oraz $Var(X^{i,j})$), estymuje się statystykami:

$$\bar{X}^{i,j} = \frac{1}{m} \sum_{k=1}^m x_k^{i,j} \quad (1)$$

$$i \quad (S^{i,j})_0^2 = \frac{1}{m-1} \sum_{k=1}^m (x_k^{i,j} - \bar{X}^{i,j})^2 \quad (2)$$

$\bar{X}^{i,j}$ i $(S^{i,j})_0^2$ są zatem przybliżeniami, ocenami (opartymi na posiadanych danych) dwóch idealnych lecz nieznanymi parametrów: $E(X^{i,j})$ i $Var(X^{i,j})$. Oczywiście, im większy rozmiar

próby, tym lepiej $\bar{X}^{i,j}$ i $(S^{i,j})_0^2$ przybliżają wartości $E(X^{i,j})$ i $Var(X^{i,j})$. Niech $\varepsilon^{i,j} = E(X^{i,j})$ oraz $v^{i,j} = Var(X^{i,j})$. Wartości te posłużą do określenia poziomu zaufania podmiotu a_i dla a_j .

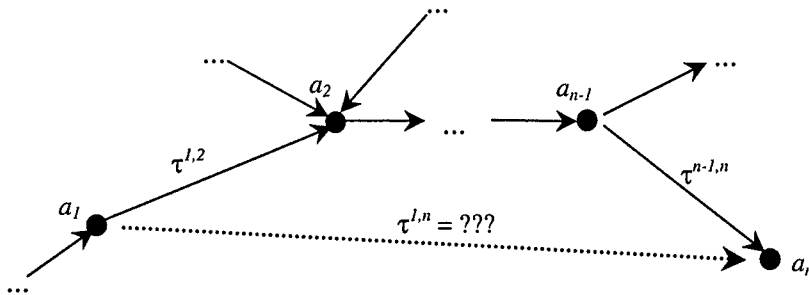
Definicja 1. Poziomem zaufania $\tau^{i,j}$ pomiędzy węzłami grafu (podmiotami) a_i oraz a_j jest para: $\tau^{i,j} = (\varepsilon^{i,j}, v^{i,j})$.

Podkreślmy, że prawdziwa wartość $\tau^{i,j}$ pozostaje nieznana. Nigdy nie posiadamy pełnej wiedzy o zmiennej losowej $X^{i,j}$. To, czym dysponujemy jest *estymowanym poziomem zaufania* $\bar{\tau}^{i,j} = (\bar{X}^{i,j}, (S^{i,j})^2)$. W dalszych obliczeniach mówiąc o $\tau^{i,j}$ będziemy wektor ten zastępowali przez $\bar{\tau}^{i,j}$. Tak zdefiniowany poziom zaufania niesie informację nie tylko o oczekiwanej ocenie kontraktu, który zawarty będzie w przyszłości ($\varepsilon^{i,j} \approx \bar{X}^{i,j}$), ale także o poziomie zmienności i niestabilności usług świadczonych przez a_j ($v^{i,j} \approx (S^{i,j})_0^2$). Pozwala to podmiotom – w podejmowanych decyzjach – uwzględniać także niepewność co do zachowań agenta, z którym współpracują. Krawędzi pomiędzy węzłami a_i oraz a_j grafu zaufania przyporządkowana jest zatem wartość $\tau^{i,j}$.

Fundamentalnym pytaniem, na jakie staramy się odpowiedzieć, jest kwestia sposobu, w jaki skończonej ścieżce w grafie można optymalnie przyporządkować poziom zaufania.

2.2. Poziom zaufania dla ścieżki w grafie

Rozważmy skierowaną ścieżkę pomiędzy węzłami a_1 oraz a_n . Przyjmijmy, że składa się ona z węzłów: a_1, a_2, \dots, a_n . Zadaniem naszym jest przyporządkować poziom zaufania podmiotu a_1 dla a_n : $\tau^{1,n} = (\varepsilon^{1,n}, v^{1,n})$ (oznaczony na rys.1 linią przerywaną).



Rys 1. Poziom zaufania pomiędzy węzłami a_1 oraz a_n

Gałęzi od a_1 do a_n przyporządkowujemy produkt $Y = X^{1,2} X^{2,3} \dots X^{n-1,n}$. Zgodnie z rozpatrywanym modelem Y jest zmienną losową. $X^{1,2}, X^{2,3}, \dots, X^{n-1,n}$ są niezależne, stąd:

$$\varepsilon^{1,n} = E(Y) = E\left(\prod_{k=1}^{n-1} X^{k,k+1}\right) = \prod_{k=1}^{n-1} E(X^{k,k+1}) = \prod_{k=1}^{n-1} \varepsilon^{k,k+1} \approx \prod_{k=1}^{n-1} \bar{X}^{k,k+1} \quad (3)$$

oraz:

$$\begin{aligned}
v^{1,n} &= \text{Var}(Y) = E(Y^2) - (E(Y))^2 = E\left(\prod_{k=1}^{n-1} (X^{k,k+1})^2\right) - \left(E\left(\prod_{k=1}^{n-1} X^{k,k+1}\right)\right)^2 = \\
&= \prod_{k=1}^{n-1} E\left((X^{k,k+1})^2\right) - \prod_{k=1}^{n-1} (E(X^{k,k+1}))^2 = \\
&= \prod_{k=1}^{n-1} (\text{Var}(X^{k,k+1}) + (E(X^{k,k+1}))^2) - \prod_{k=1}^{n-1} (E(X^{k,k+1}))^2 = \\
&= \prod_{k=1}^{n-1} (v^{k,k+1} + (\varepsilon^{k,k+1})^2) - \prod_{k=1}^{n-1} (\varepsilon^{k,k+1})^2 \approx \prod_{k=1}^{n-1} ((S^{k,k+1})_0^2 + (\bar{X}^{k,k+1})^2) - \prod_{k=1}^{n-1} (\bar{X}^{k,k+1})^2 \quad (4)
\end{aligned}$$

Łącząc (3) i (4), dla dowolnej pary komunikujących się ze sobą węzłów a_i oraz a_n grafu zaufania, przyjmujemy $\tau^{i,n} = (\varepsilon^{i,n}, v^{i,n})$.

2.3. Porządkowanie poziomów zaufania

Podstawowym problemem w porządkowaniu poziomów zaufania jest fakt, że są one wartościami wektorowymi. Jako pierwsze rozwiązanie proponujemy porządek leksykograficzny – dla danych $\tau^{a,b}$ i $\tau^{a,c}$ przyjmujemy:

$$\tau^{a,b} < \tau^{a,c} \Leftrightarrow \begin{cases} \varepsilon^{a,b} < \varepsilon^{a,c} \\ \text{lub} \\ \varepsilon^{a,b} = \varepsilon^{a,c} \wedge v^{a,b} > v^{a,c} \end{cases} \quad (5)$$

Porządek leksykograficzny zakłada, że podmiot o wyższej wartości średniej ε jest bardziej wiarygodny, dla równych wartości ε , skłonni jesteśmy wierzyć bardziej podmiotowi, którego zachowanie w przeszłości było bardziej stabilne (miało mniejszą wariancję v). Powyższe podejście nie jest jedynym. Rozważano także zastosowanie częściowych porządków stochastycznych [11] dla dystrybuant empirycznych zmiennych X (z powodu ograniczonej objętości publikacji tę część rozważań pominięto).

2.4. Łączny poziom zaufania

Założmy, że a_n należy do pewnej grupy (pracowników danej instytucji, procesów obsługiwanych i uwierzytelnianych przez danego użytkownika, etc.). Przypuśćmy ponadto, że a_i posiada już pewne doświadczenie w kontaktach z jej członkami (tj. istnieją bezpośrednie połączenia pomiędzy a_i oraz pewnymi węzłami należącymi do rozpatrywanej grupy).

Definicja 2. Niech $G = \{a_{g1}, a_{g2}, \dots, a_{gl}\}$ będzie zbiorem l podmiotów należących do tej samej grupy co a_n i takich, że w grafie zaufania istnieje połączenie pomiędzy a_i a każdym z podmiotów należących do G (co oznacza, że a_i zna pewnych członków grupy). Wartość:

$$\tau_{group}^{1,n} = \left(\frac{1}{l} \sum_{k=1}^l \varepsilon^{1,gk}, \frac{1}{l} \sum_{k=1}^l v^{1,gk} \right) = (\varepsilon_{group}^{1,n}, v_{group}^{1,n}) \quad (6)$$

nazywać będziemy *zaufaniem grupowym* pomiędzy a_l oraz a_n .

Zaufanie grupowe jest średnim poziomem zaufania, jakim a_l obdarza członków grupy, do której należy a_n (jest to użyteczna informacja, pod warunkiem, że zachowanie się a_n będzie podobne do zachowania się pozostałych członków grupy, co wydaje się być naturalne).

Przyjmijmy obecnie, że a_l posiada pewną liczbę *wysoce wiarygodnych* podmiotów. Rozsądnym jest przyjąć, że opinie ich będzie on uwzględniał w sposób szczególny. a_l sam może w dowolny sposób określić swój zbiór podmiotów *wysoce wiarygodnych*, można jednak również przyjąć, że za wysoce wiarygodny uznajemy każdy węzeł a_i , którego a_l darzy zaufaniem $\tau^{1,i}$ większym (patrz: 4.4) niż ustalona wartość progowa np. $\tau_t = (0.95, 0)$.

Definicja 3. Niech $R = \{a_{r1}, a_{r2}, \dots, a_{rh}\}$ będzie zbiorem h wysoce wiarygodnych podmiotów takich, dla których w grafie zaufania istnieje krawędź łącząca a_l z każdym z podmiotów z R a także krawędź pomiędzy każdym z podmiotów z R a węzłem a_n (oznacza to, że pewne wysoce wiarygodne w oczach a_l węzły bezpośrednio znają a_n). Wartość

$$\tau_{reputation}^{1,n} = \left(\frac{1}{h} \sum_{k=1}^h \varepsilon^{1,rk}, \frac{1}{h} \sum_{k=1}^h v^{1,rk} \right) = (\varepsilon_{reputation}^{1,n}, v_{reputation}^{1,n}) \quad (7)$$

nazywać będziemy *zaufaniem wynikającym z reputacji* pomiędzy podmiotami a_l oraz a_n (zaufanie wynikające z reputacji zdefiniowane jest dla ustalonego podzbioru wysoce wiarygodnych podmiotów). Podobnie jak poprzednio, *zaufanie wynikające z reputacji* pomiędzy podmiotami a_l oraz a_n jest "przeciętną opinią" o a_n podmiotów, których zdanie a_l ceni sobie w sposób szczególny. Określiwszy poziom zaufania, zaufanie grupowe i zaufanie wynikające z reputacji, podmiot a_l może posłużyć się nimi dla oszacowania globalnego poziomu zaufania względem a_n .

Definicja 4. Za *łączny poziom zaufania* $T^{1,n}$ podmiotu a_l do a_n przyjmuje się wartość:

$$\begin{aligned} T^{1,n} &= w_t^{1,n} \tau^{1,n} + w_g^{1,n} \tau_{group}^{1,n} + w_r^{1,n} \tau_{reputation}^{1,n} = \\ &= \left(w_t^{1,n} \varepsilon^{1,n} + w_g^{1,n} \varepsilon_{group}^{1,n} + w_r^{1,n} \varepsilon_{reputation}^{1,n}, w_t^{1,n} v^{1,n} + w_g^{1,n} v_{group}^{1,n} + w_r^{1,n} v_{reputation}^{1,n} \right) \end{aligned} \quad (8)$$

gdzie $w_t^{1,n}$, $w_g^{1,n}$, $w_r^{1,n}$ są wagami przypisanymi podmiotowi a_n przez a_l , tj.:

$$w_t^{1,n}, w_g^{1,n}, w_r^{1,n} \in [0,1] \text{ and } w_t + w_g + w_r = 1$$

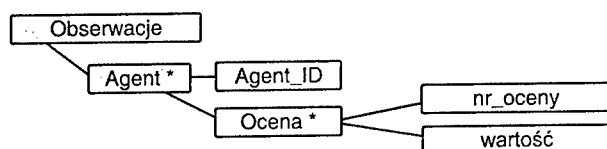
$w_g^{1,n}$ i $w_r^{1,n}$ będziemy nazywać odpowiednio wagą grupową i wagą reputacji. Zauważmy, że definicja 4. dopuszcza aby jedna (lub dwie) spośród wag były zerami. Podmiot może zatem w sposób autonomiczny przypisywać wagi wyrażając w ten sposób swoją subiektywną ocenę istotności składowych łącznego poziomu zaufania. Zauważmy też, że łączny

poziom zaufania nie będzie wykorzystywany do znakowania krawędzi grafu – jego rolą jest pomóc w przypisaniu początkowej wartości zaufania względem nowych podmiotów na podstawie przesłanek ogólniejszych niż wyłącznie τ . Łączny poziom zaufania przypisywany jest w sposób subiektywny, nie wynika on z rozważań probabilistycznych opisanych powyżej.

Przykład. Niech $\tau^{1,2} = (0.8, 0.1)$; $w_i^{1,2} = 0.6$; $w_g^{1,2} = 0.1$; $w_r^{1,2} = 0.3$; $\tau_{group}^{1,2} = (0.75, 0.05)$ i $\tau_{reputation}^{1,2} = (0.4, 0.07)$. Na podstawie definicji 4 $T^{1,2} = (0.675, 0.086)$. Można zauważyć, że stosunkowo niski poziom zaufania deklarowany przez wysoce wiarygodnych współużytkowników a_i obniża łączny poziom zaufania. Gdyby opinia wiarygodnych podmiotów była – w tym konkretnym przypadku – dla a_i bez znaczenia, wówczas przyjmując $w_i^{1,2} = 0.7$; $w_g^{1,2} = 0.3$; $w_r^{1,2} = 0.0$, otrzymalibyśmy wartość $T^{1,2} = (0.785, 0.085)$. Z kolei nie posiadając żadnych informacji o grupie, do której należy oceniany podmiot, tzn. przyjmując $w_i^{1,2} = 1$ otrzymalibyśmy $T^{1,2} = \tau^{1,2}$.

3. WYKORZYSTANIE MODELU W SYSTEMACH WEBOWYCH

Prezentowana metoda wykorzystuje, jako podstawowy typ danych, zbiory ocen przypisywanych przez agentów. Dla dalszych eksperymentów zakłada się, że będą one przechowywane w postaci plików XML (nazywanych *profilami* zaufania) i wymieniane pomiędzy agentami. Rys. 2 przedstawia drzewo odpowiedniego dokumentu XML (Element *Agent* zawiera oceny (opinia) nt. danego podmiotu działającego w sieci).



Rys. 2. Struktura pliku XML przechowującego wyniki obserwacji.
Elementy oznaczone (*) mogą być wielokrotne.

W czasie szacowania poziomu zaufania agent może zażądać od innego podmiotu zbioru obserwacji, a następnie wykorzystać otrzymane dane w obliczeniach. Dla zapewnienia poufności i autentyczności danych pliki XML powinny być kryptograficznie podpisywane i szyfrowane (z wykorzystaniem standardów XML Signatures and XML Encryption).

W proponowanym podejściu każdy z agentów może prowadzić własną politykę zaufania, na którą składają się następujące komponenty:

1. Metoda przyznawania ocen. Ta część polityki zależy od zastosowania i nie stanowi bezpośrednio części prezentowanego formalizmu. Warto jednak zaznaczyć, iż możliwe jest także wprowadzenie ocen opisowych, z których mogą korzystać użytkownicy (ludzie) – np. definiując *całkowite zaufanie* jako wartość oceny z przedziału $[0.98, 1]$.
2. Metoda ustalania wag niezbędnych do obliczenia łącznego poziomu zaufania (def.4). W ogólnym przypadku każda z grup rozpoznawanych przez agenta może otrzymać

własną wagę, w zależności od posiadanej wiedzy, dotyczącej danej grupy. Podobnie, reputacja zaufanych podmiotów może zmieniać się w czasie (np. w zależności od jakości ich ocen) itd. Zakłada się, że każdy z agentów kształtuje swoją politykę niezależnie.

3. Ponieważ w grafie zaufanie w sposób nieunikniony mamy do czynienia z wielokrotnymi ścieżkami między daną parą węzłów, należy określić jak w takiej sytuacji będzie szacowany poziom zaufania. Najprostsze podejście zakłada uwzględnienie ścieżki z najwyższym poziomem zaufania, aczkolwiek często stosowane jest także obliczanie średniej ważonej dla poziomów zaufania wielokrotnych ścieżek.

Niezależnie od powyższego, prezentowany model zaufania będzie dalej rozwijany, ze względu na otwarte wciąż interesujące problemy badawcze (nb. charakterystyczne także dla innych, opracowanych dotąd, rozwiązań). Najważniejsze obejmują:

1. Ziarnistość zaufania (*trust granularity*). Zaufanie wyrażane wobec podmiotu może zmieniać się w zależności od kontekstu. W szczególności może być różne np. dla różnych usług sieciowych świadczonych przez agenta. Dodatkowo rozróżniać możemy między zaufaniem wobec zachowania agenta oraz zaufaniem względem sformułowanych przez niego ocen innych podmiotów. Aby uwzględnić te fakty należy opracować system oznaczania ocen (a więc i wyliczonych poziomów zaufania) etykietami pozwalającymi rozróżnić kontekst wystawianej oceny.
2. Wpływ ilości obserwacji. Wydaje się oczywistym, że poziom zaufania wyznaczony na podstawie dużej ilości obserwacji (ocen) jest bardziej wiarygodny od wyliczonego na podstawie np. kilku. W proponowanym podejściu profile zaufania agentów przechowują wszystkie informacje niezbędne do uwzględnienia tego faktu.
3. Odporność na ataki, rozumiane jako obecność w sieci zaufania agentów celowo fałszujących wartości przekazywanych ocen. W tym obszarze badawczym istnieje stosunkowo niewiele opracowanych modeli i rozwiązań, głównie dotyczących modeli zaufania wykorzystywanych w systemach zarządzania kluczami (jak [9] dla X.500).

5. ZAKOŃCZENIE

Zaproponowany model spełnia wymagania sformułowane w rozdz.1 pozostając jednocześnie otwartym na dalsze modyfikacje – zakładając że podmioty (agenci) działające w sieci wymieniać będą przechowywane w profilach zaufania wektory ocen dotyczących innych podmiotów, można rozważyć (jeśli zajdzie taka potrzeba) zastosowanie do obliczania poziomów zaufania metod innych niż opisane w rozdz.2. Podobnie nie stanowi problemu skojarzenie z wektorami ocen informacji o kontekście, co pozwoliłoby na realizację postulatu ziarnistości zaufania. Ścisła probabilistyczna interpretacja pozwala także na wykorzystanie opisanej metryki w kontekście oceny niezawodności (jakości usług) komponentów systemu webowego.

Badanie modeli zaufania jest obecnie dynamicznie rozwijającą się dziedziną, związana głównie z gwałtownym rozwojem systemów webowych a w szczególności standardów opisu i przetwarzania wiedzy w takich systemach (jak XML, RDF, DAML...)[2]. Wypracowanie metod oceny jakości udostępnianej przez działające w sieci podmioty informacji staje się koniecznością.

BIBLIOGRAFIA

- [1] Adamic L., "The Small World Web". Proceedings of ECDL, pp. 443-452, 1999.
- [2] Berners-Lee T. *Semantic Web Road Map*. World wide Web Consortium (W3C) Design Issues, October 1998.
- [3] Blaze M. et. al. The role of trust management in distributed systems security. *Secure Internet Programming: Security Issues for Mobile and Distributed Objects*, pp. 185-210. Springer Verlag, 1999.
- [4] Buskens V., The social structure of trust, *Social Networks* 20 pp. 265-289, Elsevier 1998.
- [5] Dumbill E., "XML Watch: Finding friends with XML and RDF." IBM Developer Works, at: <http://www-106.ibm.com/developerworks/xml/library/xfoaf>, June 2002.
- [6] Golbeck J., Hendler J., Parsia B., Trust Networks on the Semantic Web, Proceedings of Cooperative Intelligent Agents 2003, Helsinki, Finland.
- [7] Yao-Hua Tan, Walter Thoen, Formal aspects of a generic model of trust for electronic commerce Decision Support Systems 33 pp.233-246, Elsevier, 2002.
- [8] Reagle J. M. Trust in a cryptographic economy and digital security deposits: Protocols and policies. Master of Science Thesis, Department of Technology and Policy, MIT, 1996.
- [9] Twigg A., Dimmock N, Attack-Resistance of Computational Trust Models, WETICE03, Linz, Austria, IEEE Press, 2003, pp.275-280.
- [10] Venkatraman M. et. al., Trust and Reputation Management in a Small-World Network, Proceedings of Fourth International Conference on MultiAgent Systems, pp. 449-450, 2000
- [11] Stoyan D., Comparison Methods for Queues and Other Stochastic Models, J.Wiley & Sons, Berlin, 1983.

TRUST MODEL FOR WEB SYSTEMS**Summary**

A probabilistic approach for trust assessing and representation is proposed. Our approach postulates a simple representation of trust primitives (sets of ratings) which are then used by more sophisticated methods. Proposed solution is flexible and allows each of the network agents to conduct his own trust. The model contains also a method of estimating group and reputation trust.

Jacek Stefański

Katedra Systemów i Sieci Radiokomunikacyjnych, Politechnika Gdańska

TRENDY ROZWOJOWE TECHNOLOGII RADIA PROGRAMOWALNEGO

Streszczenie

W referacie omówiono koncepcję technologii radia programowalnego, która w niedalekiej przyszłości zdominuje rynek usług radiokomunikacyjnych. Przedstawiono główne tendencje rozwojowe radia programowalnego oraz ograniczenia wynikające z praktycznej realizacji terminali ruchomych w tej technologii. Zaproponowano również infrastrukturę teleinformatyczną dla sieci radiokomunikacyjnej, wykorzystującej technologię radia programowalnego.

1. WPROWADZENIE

W dobie gwałtownego rozwoju systemów radiowej transmisji danych, ukierunkowanych na świadczenie usług multimedialnych, istnieje paląca potrzeba wprowadzenia do powszechnego użytku nowej jakościowo technologii radia programowalnego SDR (ang. *Software Defined Radio*) [1, 2]. Zapewni ona otwartość i uniwersalność już istniejących i nowych systemów radiokomunikacyjnych. Technologia radia programowalnego umożliwi współpracę terminala ruchomego z wieloma systemami, działającymi w różnych standardach oraz implementację nowych usług, bez konieczności wymiany wyposażenia terminala. Będzie to możliwe dzięki zastąpieniu realizowanych współcześnie procesów przetwarzania sygnału radiowego, wykonywanych oddzielnie przez poszczególne człony toru nadawczo-odbiorczego, specjalistycznym oprogramowaniem na bazie uniwersalnej platformy sprzętowej. Oznacza to, że postać sprzętową współczesnego terminala ruchomego zastąpi w dominującym stopniu oprogramowanie rezydujące w nowym terminalu.

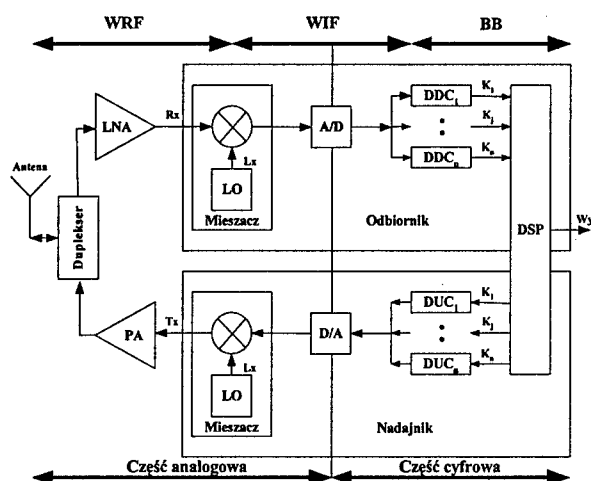
Terminal ruchomy wykonany w technologii radia programowalnego będzie więc uniwersalnym urządzeniem, w którym realizowane sprzętowo funkcje toru nadawczo-odbiorczego będą sprowadzone do niezbędnego minimum [3]. Natomiast jądrem tego toru będzie procesor sygnałowy, współpracujący z szerokopasmowymi przetwornikami analogowo-cyfrowymi oraz cyfrowo-analogowymi [4]. Procesor sygnałowy będzie mieć możliwość zdalnej wymiany oprogramowania (najlepiej bez udziału użytkownika), które wyznaczy funkcje terminala ruchomego, związane z jego współdziałaniem w danym systemie i zapewniające realizację określonych usług.

2. ARCHITEKTURA RADIA PROGRAMOWALNEGO

Opracowanie możliwie uniwersalnej platformy sprzętowej dla potrzeb technologii radia programowalnego wiąże się z zapewnieniem programowego sterowania następującymi charakterystykami i/lub parametrami toru nadawczo-odbiorczego:

- zakresem przestrajania nadajnika i odbiornika w paśmie w.cz.,
- wyborem częstotliwości nośnej kanału roboczego,
- szerokością pasma sygnału nadawanego,
- selektywnością odbiornika,
- rodzajem kompresji/dekompresji informacji źródłowej,
- rodzajem modulacji oraz metodą detekcji sygnału,
- rodzajem i stopniem kodowego zabezpieczenia przed błędami wprowadzanymi przez kanał,
- maksymalną przepływnością w kanale radiowym,
- obsługą strumieni informacji podlegających i nie podlegających ograniczeniom czasu rzeczywistego,
- częstotliwością pośrednią w odbiorniku i nadajniku,
- mocą sygnału nadawanego,
- częstotliwością próbkowania przetworników analogowo-cyfrowych i cyfrowo-analogowych,
- charakterystykami filtrów toru nadawczo-odbiorczego.

Schemat blokowy architektury radia programowalnego, spełniający powyższe postulaty, został przedstawiony na rys.1 [5]. Sygnały wysyłane z nadajnika programowalnego przez kanał radiowy za pomocą fali elektromagnetycznej trafiają do szerokopasmowej anteny odbiornika programowalnego, skąd po wzmocnieniu w niskoszumnym wzmacniaczu wysokiej częstotliwości LNA są podawane następnie, poprzez duplekser, na wejście mieszacza. Duplekser służy do niezależnego odbioru i nadawania sygnałów w paśmie wysokiej częstotliwości. W mieszaczu dokonuje się transformacja szerokopasmowego sygnału odebranego w paśmie wysokiej częstotliwości WRF do pasma pośredniej częstotliwości WIF, w którym możliwe staje się przekształcenie sygnałów analogowych w ich reprezentację cyfrową, za pomocą szybkich przetworników analogowo-cyfrowych A/D. Pomocniczy sygnał harmoniczny z zadanego zakresu częstotliwości L_x , dostarczany do bloku mieszacza, pochodzi z oscylatora lokalnego LO, wykonanego w postaci syntezy częstotliwości. W tradycyjnej technologii radiowej stopień przemiany pośredniej częstotliwości zapewnia wyselekcjonowanie odpowiedniego kanału częstotliwościowego. Natomiast w technologii radia programowalnego za selekcję odpowiedniego kanału częstotliwościowego (tj. pobieranie próbek sygnału na wyjściu przetwornika A/D, związanych z określonym kanałem) odpowiada tzw. blok cyfrowej przemiany częstotliwości DDC. W skład bloku DDC wchodzi trzy główne człony: cyfrowy oscylator lokalny, cyfrowy mieszacz zespolony oraz cyfrowy filtr dolnopasmowy o skończonej odpowiedzi impulsowej. Cyfrowy oscylator lokalny dostarcza do mieszacza próbki sygnału zespolonego (próbki składowych sinusoidalnych i kosinusoidalnych). Jest on wykonany w postaci bezpośredniego cyfrowego syntezy częstotliwości. Cyfrowy mieszacz zespolony jest odpowiedzialny za wymnożenie próbek sygnału, pochodzących z przetwornika A/D, przez próbki składowych sinusoidalnych i kosinusoidalnych, pojawiających się na wyjściu cyfrowego oscylatora lokalnego. Dokonuje on więc transformacji sygnału z pasma pośredniej częstotliwości do pasma podstawowego BB.



Rys.1. Architektura radia programowalnego

Oznaczenia: A/D – przetwornik analogowo-cyfrowy (ang. *Analog to Digital converter*),
 BB – praca w paśmie podstawowym (ang. *BaseBand*),
 D/A – przetwornik cyfrowo-analogowy (ang. *Digital to Analog converter*),
 DDC – cyfrowa przemiana częstotliwości „w dół” (ang. *Digital Down Converter*),
 DSP – człon cyfrowego przetwarzania sygnału, procesor sygnałowy (ang. *Digital Signal Processor*),
 DUC – cyfrowa przemiana częstotliwości „w górę” (ang. *Digital Up Converter*),
 LNA – niskoszumny wzmacniacz wysokiej częstotliwości (ang. *Low Noise Amplifier*),
 LO – oscylator lokalny (ang. *Local Oscillator*),
 PA – wzmacniacz mocy wysokiej częstotliwości (ang. *Power Amplifier*),
 WIF – szerokopasmowy człon pośredniej częstotliwości (ang. *Wideband Intermediate Frequency*),
 WRF – szerokopasmowy człon wysokiej częstotliwości (ang. *Wideband Radio Frequency*).

Cyfrowy filtr dolnoprzepustowy dokonuje selekcji sygnału o wybranej częstotliwości nośnej i o określonym paśmie. Blok DDC stanowi zatem kolejny stopień przemiany częstotliwości w odbiorniku, realizowanej w sposób cyfrowy. Po stronie nadawczej znajdują się komplementarne bloki funkcjonalne do wyżej omówionych, czyli tzw. cyfrowa przemiana częstotliwości DUC, przetwornik cyfrowo-analogowy D/A, dostarczający szerokopasmowy sygnał cyfrowy z pasma pośredniej częstotliwości do mieszacza w członie nadajnika oraz wzmacniacz mocy PA, zapewniający odpowiednią moc sygnału nadawanego na zaciskach anteny. Człon cyfrowego przetwarzania sygnałów DSP w radiu programowalnym realizuje funkcje toru nadawczo-odbiorczego cyfrowego systemu radio-komunikacyjnego w paśmie podstawowym, tzn. kodowanie i dekodowanie źródłowe, kodowanie i dekodowanie kanałowe, szyfrację i deszyfrację, modulację i demodulację, względnie dodatkowo rozpraszanie i skupianie widma sygnałów w zależności od stosowanej techniki wielodostępu.

3. OGRANICZENIA TECHNOLOGICZNE

Omówiona powyżej architektura radia programowalnego, która została przyjęta za punkt wyjścia przy budowie platformy sprzętowej SDR, napotyka jednak ciągle na problemy technologiczne, związane z implementacją poszczególnych członów toru nadawczo-

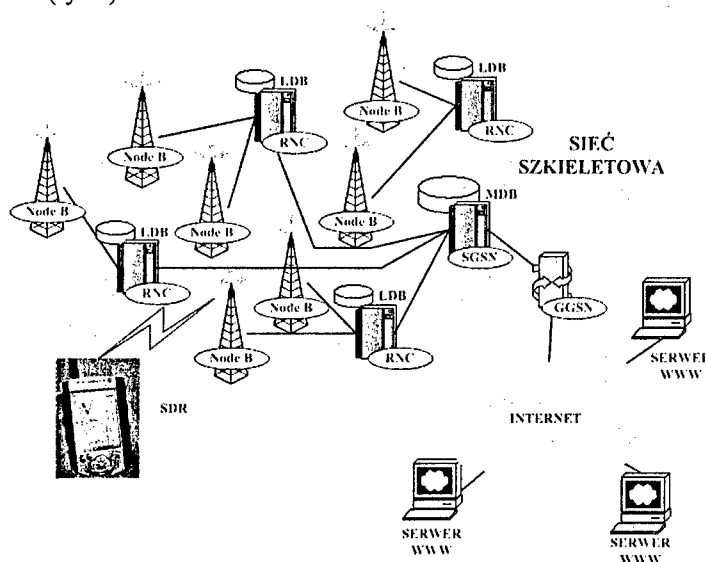
odbiorczego terminala ruchomego w czasie rzeczywistym. Pierwszą napotkaną barierą jest wytworzenie szybkich przetworników A/D i D/A, zapewniających odpowiednią rozdzielczość przetwarzania [6]. Szacuje się, że rozdzielczość ta powinna być rzędu 16 bitów, co zapewni dynamikę przetwarzania na poziomie 100 dB. Z trendów rozwojowych tej gałęzi produkcji układów scalonych wynika, że zwiększanie częstotliwości próbkowania w konstruowanych przetwornikach nie idzie w parze ze wzrostem rozdzielczości. Z analizy parametrów elektrycznych przetworników A/D (wynikającej z teorii nieoznaczoności Heisenberg'a) wynika, że granicę technologiczną stanowi wyprodukowanie przetwornika pracującego z częstotliwością próbkowania 1 GHz i rozdzielczością 20 bitów, co zapewnia dynamikę przetwarzania na poziomie 120 dB. Dalsze polepszenie parametrów elektrycznych przetworników może być realizowalne jedynie po znacznym obniżeniu temperatury otoczenia pracy przetwornika (do pojedynczych Kelvinów). Oczywiście spełnienie tego ostatniego postulatu trudno na dzień dzisiejszy wyobrazić sobie w sprzęcie przenośnym. Kłopoty technologiczne z wytworzeniem przetworników D/A dla potrzeb technologii radia programowalnego są związane z zapewnieniem odpowiednio wysokiej liniowości przetwarzania, integracji filtru wyjściowego w jednym układzie scalonym oraz skutecznej izolacji sygnałów taktujących pracę przetwornika od jego wyjścia analogowego w celu zminimalizowania zakłóceń.

Oddzielnym problemem jest wytworzenie szybkich procesorów sygnałowych, mogących sprostać wymaganej szybkości cyfrowej obróbki sygnałów w czasie rzeczywistym. Jak wynika z licznych badań symulacyjnych, prowadzonych dla potrzeb systemu UMTS (ang. *Universal Mobile Telecommunications Systems*), wydajność obliczeniowa procesorów sygnałowych do implementacji terminali w technologii radia programowalnego powinna sięgać dziesiątek miliardów instrukcji/s, przy zadowalającej dokładności obliczeń. Na dzień dzisiejszy potentaci w produkcji procesorów sygnałowych oferują szybkość obliczeniową swoich produktów o przynajmniej jeden rząd mniejszą od wymaganej. Z analizy rynku układów scalonych wynika, że na zwiększenie wydajności obliczeniowej procesorów sygnałowych o miliard instrukcji/s potrzeba około jednego roku. Dodatkowym utrudnieniem w zapewnieniu odpowiednich procesorów sygnałowych do przyszłych zastosowań w terminalach ruchomych, podobnie jak to ma miejsce przy wytwarzaniu przetworników A/D i D/A, jest korelacja pomiędzy wzrostem szybkości przetwarzania a poborem mocy, tzn. im wyższa jest wydajność obliczeniowa procesora sygnałowego, tym pobór mocy ze źródła jest większy, co nie jest korzystne z punktu widzenia zastosowań w sprzęcie ruchomym. W celu obniżenia wymaganej szybkości przetwarzania sygnałów w DSP proponuje się wyodrębnienie cyfrowej przemiany częstotliwości DDC i DUC na zewnątrz procesora sygnałowego w postaci programowalnych matryc FPGA (ang. *Field Programmable Gate Array*) (patrz rys.1). Pomimo tego, dostępne obecnie procesory sygnałowe nie będą mogły sprostać rygorystycznemu rynkowi radia programowalnego. W związku z tym proponuje się systemy wieloprocessorowe o odpowiedniej mocy obliczeniowej, z których buduje się na razie jedynie stacje bazowe w technologii radia programowalnego. Stanowią one poligon doświadczalny nowej technologii, a uzyskane tą drogą rezultaty będzie można w przyszłości z powodzeniem przenieść do terminali przenośnych.

4. INFRASTRUKTURA TELEINFORMATYCZNA

Wprowadzenie technologii radia programowalnego będzie się wiązało z koniecznością przesyłania i przechowywania znacznych ilości danych, przede wszystkim wspomnianego wcześniej oprogramowania procesorów sygnałowych, które będzie decydować o funkcjonalności terminala ruchomego. Pod tym względem, dotychczasowa infrastruktura teleinfor-

matyczna sieci komórkowych może się okazać niewystarczająca do sprawnej obsługi tej technologii. Poza tym, znaczne rozproszenie geograficzne wyodrębnionych urządzeń sieci komórkowych utrudnia zbudowanie w stosunkowo krótkim czasie nowej sieci teleinformatycznej. W związku z tym sprawne wprowadzenie do użytku technologii radia programowalnego może wymusić na operatorach sieci radiokomunikacyjnych skorzystanie z zasobów zewnętrznych dostawców usług teleinformatycznych. Wstępnie proponowana architektura sieci teleinformatycznej w radiokomunikacyjnym systemie trzeciej generacji UMTS dla potrzeb technologii radia programowalnego może się przedstawiać w następujący sposób (rys.2).



Rys.2. Przykładowa architektura sieci UMTS dla potrzeb radia programowalnego.

Oznaczenia: GGSN – węzeł tranzytowy podsystemu GPRS (ang. *Gateway General packet radio service Support Node*),

LDB – lokalna baza danych (ang. *Local Data Base*),

MDB – główna baza danych (ang. *Main Data Base*),

Node B – stacja bazowa,

RNC – sterownik sieci radiowej (ang. *Radio Network Controller*),

SDR – radio programowalne (ang. *Software Defined Radio*),

SGSN – węzeł obsługujący podsystem GPRS (ang. *Serving General packet radio service Support Node*).

Komunikacja sieci szkieletowej CN (ang. *Core Network*) z zewnętrzną siecią teleinformatyczną odbywać się może za pośrednictwem węzła GGSN. Natomiast z węzłem SGSN proponuje się skojarzyć tzw. główną bazę danych MDB, w której będzie przechowywane pełne oprogramowanie dla potrzeb obsługi technologii radia programowalnego. Oprogramowanie to pochodzić będzie z poszczególnych serwerów WWW jego wytwórców. Obok tego, z każdym sterownikiem sieci radiowej RNC lub grupą współ-pracujących ze sobą takich sterowników, proponuje się skojarzyć lokalne bazy danych LDB. Bazy te będą połączone łączami stałymi z bazą główną. W każdej bazie lokalnej będą przechowywane kopie zasobów bazy głównej, każdorazowo aktualizowanych po wprowadzeniu zmian w tej bazie. Poza niewątpliwymi zaletami proponowanego rozwiązania, tzn. szybkim dostępem do aktualnego oprogramowania oraz prostotą zarządzania siecią, podejście takie stanowi

może drogie rozwiązanie, ponieważ wymaga budowy dużej liczby lokalnych baz danych w stosunkowo krótkim czasie (wprowadzenie technologii radia programowalnego wiąże się z udostępnieniem w każdym miejscu sieci radiokomunikacyjnej wymaganego oprogramowania). Dla zmniejszenia kosztów proponuje się zrezygnować z lokalnych baz danych (przynajmniej w pierwszej fazie powstawania infrastruktury teleinformatycznej) [7]. Natomiast główną bazę danych MDB można zastąpić serwerem WWW, świadczącym między innymi usługi dzierżawy pamięci dyskowej na zasadach komercyjnych. Oznacza to, że do prawidłowego funkcjonowania sieci teleinformatycznej nie będą potrzebne łącza stałe. Z uwagi na rozległość sieci szkieletowej dopuszcza się dzierżawę innych serwerów WWW, znajdujących się bliżej określonej grupy modułów RNC, np. na obszarze silnie zurbanizowanym.

5. PODSUMOWANIE

Przewiduje się, że technologia radia programowalnego będzie wprowadzana do eksploatacji w latach 2005-2010. Oprócz nowych osiągnięć technologii układów scalonych, motorem rozwoju technologii SDR w terminalach ruchomych jest tzw. ewolucyjna koncepcja rozwoju systemów radiokomunikacji komórkowej, czyli zainicjowany już proces łagodnego przejścia od systemów drugiej generacji do systemu UMTS. Ograniczenia związane z wprowadzaniem technologii radia programowalnego nie dotyczą tylko aspektów technologicznych, ale również są związane z niedostatecznie rozwiniętym segmentem teleinformatycznym sieci radiokomunikacyjnych poszczególnych operatorów.

BIBLIOGRAFIA

- [1] Buracchini E.: *The Software Radio Concept*. IEEE Communications Magazine, s. 138-143, September 2000.
- [2] Mitola J. III: *Software Radio Architecture. Object-Oriented Approaches to Wireless Systems Engineering*. John Wiley & Sons, 2000.
- [3] Walewski K.: *Technologia terminala ruchomego w systemach 3G*. Materiały konferencyjne Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji, s. 481-484, Wrocław, 25-27 czerwiec 2003.
- [4] Stefański J., Gajewski S., Marczak A.: *Radio rekonfigurowalne programowo w systemie UMTS*. Elektronik, nr 11, s. 49-53, listopad 2001.
- [5] Efstathiou D., Fridman J., Zvonar Z.: *Recent Developments in Enabling Technologies for Software Defined Radio*. IEEE Communications Magazine, s. 112-117, August 1999.
- [6] Harada H., Prasad R.: *Simulation and Software Radio for Mobile Communications*, Artech House, 2002.
- [7] Katulski R., Stefański J.: *Uwarunkowania teleinformatyczne technologii radia programowalnego*. Materiały konferencyjne Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji, s. 519-522, Wrocław, 25-27 czerwiec 2003.

THE DEVELOPMENT OF SOFTWARE DEFINED RADIO TECHNOLOGIES

Summary

The concept of software defined radio has been outlined and the main trends and constraints in software defined radio were presented. The anticipated architecture of radio communication networks based on software defined radio technologies, was proposed. It seems that in the nearest future the market of radio communications services will be dominated by new technologies.

Mariusz Wrzesień

Katedra Systemów Ekspertowych i Sztucznej inteligencji,
Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie

OPTYMALIZACJA PEWNYCH FUNKCJI KLASYFIKATORA BINARNEGO W PROCESIE POZYSKIWANIA INFORMACJI Z DANYCH WIELOKATEGORYJNYCH

Streszczenie

W artykule opisano kolejny etap badań nad rozwojem pewnych funkcji klasyfikatora binarnego stosowanego do pozyskiwania informacji z danych wielokategoryjnych. Przedstawiono najważniejsze moduły programowe, oraz zastosowany pierwszy z etapów optymalizacji procesu uczenia. Zaprezentowano wyniki przeprowadzonego eksperymentu oraz wnioski i przyszłe kierunki badań.

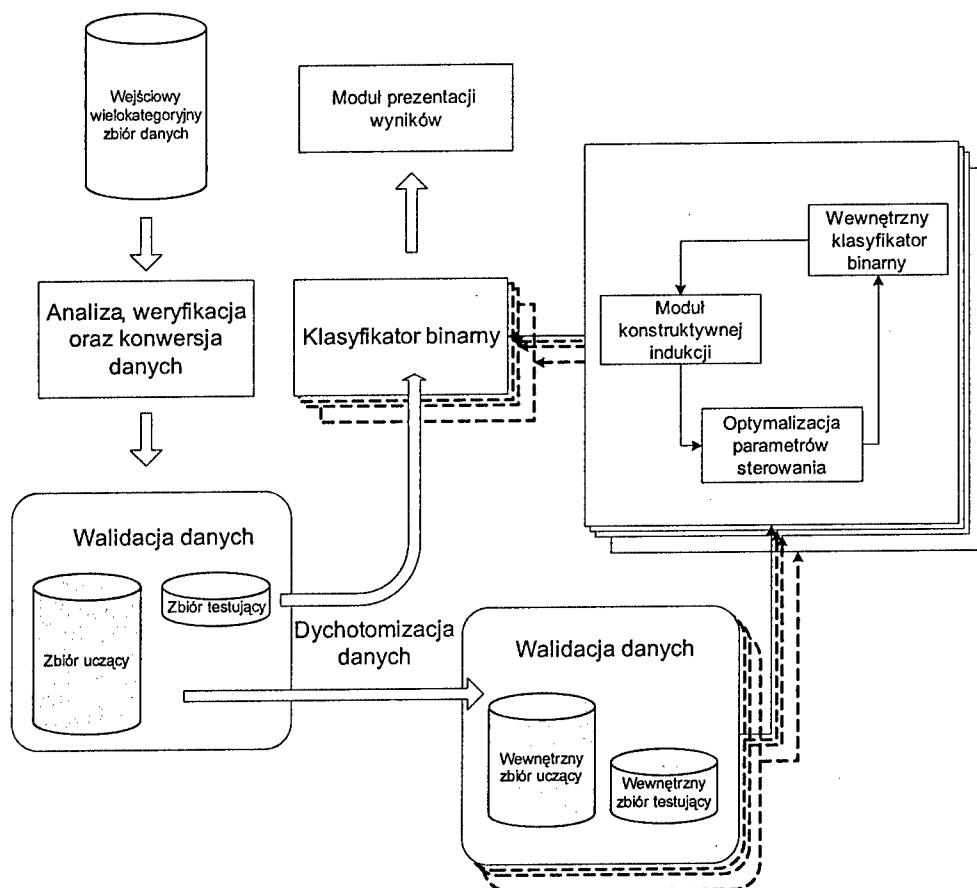
1. WSTĘP

Badania opisane w niniejszej pracy dotyczą zastosowań elementów nadzorowanego uczenia maszynowego w klasyfikacji i identyfikacji obiektów fizycznych lub/ oraz abstrakcji, np. zmian melanocytowych skóry. Osiągnięcie tego celu wymagało opracowania suity narzędzi informatycznych, umożliwiających wyszukiwanie informacji i wiedzy ukrytych w danych, przy czym – co jest dość powszechne w przypadku niektórych danych, np. medycznych – mogą być one niepewne, a często nawet sprzeczne. W ramach wspomnianych badań w Katedrze Systemów Ekspertowych i Sztucznej Inteligencji WSiIZ zostały opracowane następujące narzędzia: *AffinitySEEKER*[®] (wykorzystujący metody minimalno-odległościowe do poszukiwania podobieństwa pomiędzy identyfikowanymi obiektami) [1], *ReliefSEEKER*[®] (generujący stochastyczne sieci przekonań) [2], *PlaneSEEKER*[®] (wykorzystujący zoptymalizowane algorytmy liniowej maszyny uczącej do identyfikacji obiektów wielokategoryjnych, z zastosowaniem rekurencyjnego klasyfikatora binarnego) [3] oraz *TreeSEEKER*[®] (generujący quasi-optimalne drzewa decyzji) [4]. Dodatkowym, opracowanym narzędziem informatycznym (z poza problematyki pozyskiwania informacji i wiedzy), ułatwiającym niejako korzystanie z czterech wymienionych aplikacji, jest *ScoreSEEKER* (przygotowujący, *n*-par plików do przeprowadzenia *n*-krotnej skróśnej oceny skuteczności działania narzędzi pozyskujących wiedzę). W dalszej części artykułu zostaną omówione algorytmy zastosowane w trakcie budowy narzędzia *PlaneSEEKER*.

2. OGÓLNA CHARAKTERYSTYKA BUDOWANEGO NARZĘDZIA

PlaneSEEKER

Budowane narzędzie informatyczne ma w sobie zaimplementowany algorytm obsługi dużych baz informacyjnych (tekstowych). Moduł ten został zaprojektowany z myślą nadania opracowanemu narzędziu zdolności przetwarzania dużych tablic decyzji, zawierających ponad 3000 wielokategoryjnych przypadków. Wewnętrzne algorytmy generują model uczenia (zbiór wektorów wagowych), stosując uciążliwe obliczeniowo, oraz skrupulatne metody oceny jakości klasyfikacji i identyfikacji badanych obiektów. Z tego względu czas maszynowy wymagany do zrealizowania zadania klasyfikacji, może być nieoczekiwanie bardzo duży (rzędu kilku do kilkunastu godzin). W celu uniknięcia tej sytuacji, blok obsługi dużych baz informacyjnych sekwencyjnie wczytuje „paczki” danych do pamięci operacyjnej, tworząc zgodnie z decyzją użytkownika: (i) odzwierciedlenie źródłowych danych, z ścisłym zachowaniem proporcji liczby przypadków opisujących poszczególne kategorie, albo (ii) reprezentację klas (najbardziej „podobne” przypadki dla każdej kategorii reprezentowanej w tablicy decyzji).



Rys. 1. Schemat blokowy budowanego narzędzia zawierający najważniejsze moduły

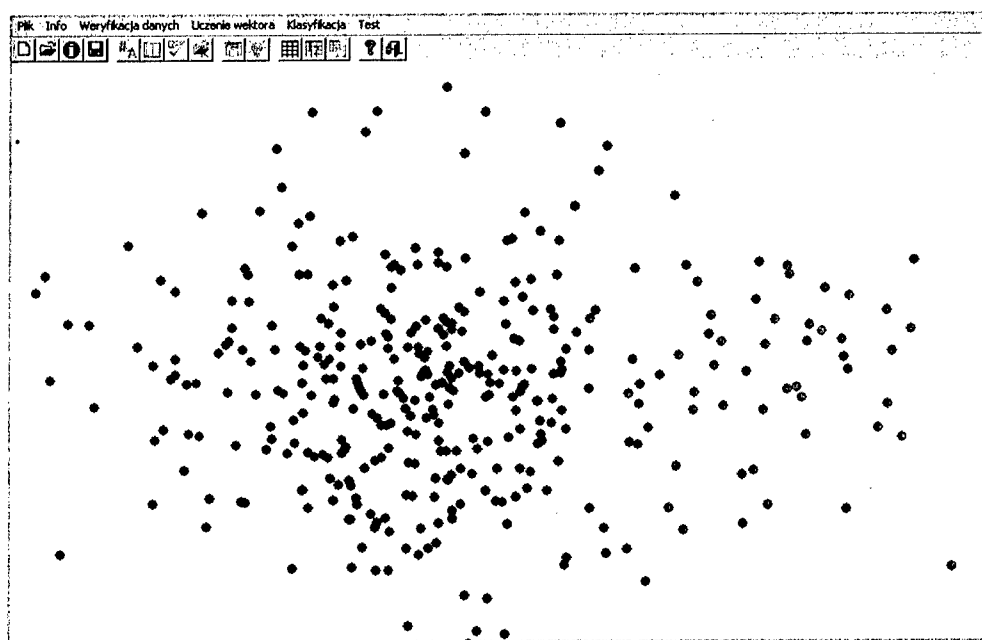
Podczas sekwencyjnego wczytywania danych źródłowych, pamięć operacyjna jest stopniowo uwalniana, zaś wersja robocza tablicy decyzji zostaje ostatecznie zapisana na dysku. Blok analizy danych źródłowych obejmuje zestaw zaimplementowanych algorytmów pobierania danych wejściowych, zawartych w tekstowych tablicach decyzji wg. Pawlaka [5], tzn. w tablicach typu 2a [6] (dowolna liczba zmiennych opisujących, jedna zmienna złożona, zlokalizowana w ostatniej kolumnie). Dodatkowo, realizowane są funkcje autokorekty danych, których celem jest wykrycie i usunięcie niezgodności oraz błędów powstałych podczas tworzenia bazy informacyjnej. Wbudowany algorytm tzw. wstępnego przetwarzania danych, umożliwia automatyczne wykrywanie błędów „literowych” w nazwach atrybutów oraz w ich wartościach (gdy są symboliczne), a także umożliwia uzupełnianie lub eliminację brakujących i błędnych wartości. Działanie omawianego bloku może być automatyczne lub pod nadzorem użytkownika.

Specyfika działania tego typu narzędzi informatycznych umożliwia pracę na ogół jedynie na danych numerycznych (ciągłych), dlatego też system został wyposażony w moduł do konwersji danych symbolicznych w numeryczne. Wg danych opisanych w dostępnej literaturze [7], sposób konwersji danych symbolicznych może mieć istotny wpływ na poprawność identyfikacji i klasyfikacji nieznanych przypadków. W budowanym narzędziu informatycznym do pozyskiwania informacji z danych wielokategoryjnych, wykorzystano specyficzne podejście do analizy informacyjnej, polegające na zastosowaniu klasyfikatora binarnego [8, 9, 10], bazującego na podziale n -kategoryjnego zbioru na Pn zbiorów dychotomicznych [7]. Koncepcja klasyfikacji wielokategoryjnej poprzez dychotomizację plików wielokategoryjnych jest stosowana również przez inne grupy badawcze, (np. [11]), jednakże z zupełnie innych powodów. W przypadku tym rozkład wielokategoryjnych obiektów w Pn dychotomicznych plikach ujmuje wyczerpująco wszystkie kombinacje przypadków o dwóch klasach. Tak więc, gdy pierwotna baza danych zawiera obiekty należące np. do 4 klas (powiedzmy, A, B, C i D), wtedy system generuje $4(4-1)/2 = 6$ dychotomicznych kombinacji, z których każda zawiera przypadki należące do dwóch kategorii. W rozpatrywanym przykładzie będzie to kombinacja klas: AB, AC, AD, BC, BD oraz CD. W procesie uczenia generowany jest dla każdego ze zbiorów dychotomicznych wektor wagowy, zdolny do klasyfikacji nieznanych przypadków w obrębie klas, na których uczył się klasyfikacji. Zbiór wektorów wagowych (ZWW) tworzy globalny model uczenia - model ten może być zastosowany w procesie klasyfikacji nieznanego (nieznanych) przypadku (przypadków). Wtedy, każdy z wektorów wagowych przypisuje nieznanemu obiektowi jedną z dwóch możliwych klas. Jeżeli rozpoznawany obiekt będzie należał np. do klasy A, to prawdziwą decyzję będą mogły podjąć tylko te klasyfikatory, które były uczone na zbiorach zawierających przypadki z klasy A (zbiory: AB, AC, AD), pozostałe klasyfikatory uczone na zbiorach nie zawierających przypadków z klasy A (zbiory: BC, BD, CD) będą generowały odpowiedzi nieprzewidywalne, najczęściej fałszywe. Klasa, która najczęściej zostanie rozpoznana przez ZWW jest przypisana nieznanemu obiektowi. Wydaje się, że poprawność klasyfikacji przypadków należących do większej od dwóch liczby kategorii, może być bardzo wysoka, gdy poprawność klasyfikacji poszczególnych (indywidualnych) klasyfikatorów binarnych nie wzbudza zastrzeżeń [12]. Na podstawie wstępnych wyników badań opisanych w [3] można wnioskować, że zaimplementowany algorytm działa poprawnie bez względu na to czy zbiory pierwotne były 3, 4, lub 5-cio kategoryjne.

Budowane narzędzie informatyczne zostało dodatkowo wyposażone w moduły umożliwiające kompleksowe prowadzenie badań. *Moduł walidacji danych* umożliwia przeprowadzenie niezależnej estymacji trafności klasyfikacji w oparciu o:

- wielokrotną skrośną walidację,
- technikę leaving-one-out oraz,
- procentowy podział zbioru danych.

Moduł wizualizacji danych źródłowych umożliwia przedstawienie przypadków zawartych w bazie informacyjnej w postaci punktów w trójwymiarowej przestrzeni rozwiązań. Szczegółowe omówienie wewnętrznych algorytmów tego modułu nie jest możliwym z uwagi na ograniczenia redakcyjne rozmiaru publikacji. Przykładowy wgląd typu 2D, uzyskany za pomocą omawianego modułu został przedstawiony na rysunku Rys. 2.



Rys. 2. Dwuwymiarowy wgląd w treść bazy zmian melanocytowych skóry P548144

System został również wyposażony w *moduł prezentacji wyników* w którym podczas badań prezentowana jest klasyfikacja (nieznanych przypadków) poprzez poszczególne klasyfikatory bazowe.

Specyfika działania klasyfikatora binarnego wymaga kilkupoziomowej optymalizacji jego pracy. W pierwszym etapie badań zwrócono uwagę na odpowiedni dobór parametrów zarządzających procesem uczenia do których należą algorytmy opisane w [13].

3. EKSPERYMENT

Budowane narzędzie oraz zastosowane w nim algorytmy zostały sprawdzone na wielu różnych bazach informacyjnych, pochodzących z Repozytorium Baz Danych dla uczenie

maszynowego [14]. W niniejszym artykule, z uwagi na narzucone ograniczenia objętość tekstu, zostały przedstawione jedynie wybrane wyniki badań m.in. dla przypadków zgromadzonych w bazie informacyjnej P548144 opisanej w [15]. Badany zbiór danych zawiera 548 przypadków, zdefiniowanych przy pomocy 13 atrybutów, oraz atrybut TDS, który jest sprawdzeniem działania konstruktywnej indukcji, rozpatrywane przypadki opisują znamiona melanocytowe skóry w czterech kategoriach. Dobór parametrów sterowania odbywał się zawsze na podstawie 70% przypadków zbioru uczącego, natomiast jakość modelu uczenia była testowana na pozostałych 30% przypadków zbioru uczącego. W tablicy 3.1 zaprezentowano wyniki przeprowadzonych badań, w kolumnie drugiej został zamieszczony błąd klasyfikacji zanotowany przy pracy klasyfikatora binarnego bez modułu optymalizacji, natomiast w kolumnie trzeciej zamieszczono błąd klasyfikacji po wykonaniu pierwszego etapu optymalizacji. Rzeczywista jakość uczenia była oceniana z zastosowaniem dziesięciokrotnej skróśnej walidacji.

Tablica 3.1

Wyniki przeprowadzonych badań

Nazwa badanej bazy	Błąd klasyfikacji [%]	Błąd klasyfikacji po I etapie optymalizacji [%]
Australian [14]	36,51	23,43
Baza zmian melanocytowych skóry P548144 [15]	32,21	10,62
Car [18]	29,20	19,80
Contraceptive Method Choice [16]	47,54	44,12
Hvote [14]	8,70	8,70
Iris Plants Database [14]	7,33	2,66
Johns Hopkins University Ionosphere database [14]	17,93	13,10
Sunb [17]	19,55	4,33

Wyniki badań zestawionych w tablicy 4.1 wskazują, że optymalizacja procesu uczenia wektorów wagowych przez kontrolowane sterowanie parametrami procesu uczenia wywiera istotny wpływ na zmniejszenie błędu klasyfikacji. Natomiast kolejny etap optymalizacji powinien doprowadzić – przynajmniej dla badanych przez nas zbiorów danych – do poziomu błędu uzyskiwanego za pomocą innych modeli uczenia, np. drzew decyzji czy sieci przekonań [19].

4. PRZYSZŁE KIERUNKI BADAŃ

W najbliższym czasie planowana jest implementacja drugiego etapu optymalizacji procesu klasyfikacji przypadków wielokategoryjnych. W etapie tym postanowiono podjąć próbę usprawnienia klasyfikatora binarnego poprzez sięgnięcie do przyczyn jego zróżnicowanej trafności klasyfikowania. Obiecującą możliwością usprawnienia działania tego typu

klasyfikatora wydają się być zastosowanie np. sieci przekonanych [19] w konstruktywnej indukcji [20]. Operacja ta w zarysie polega na dodaniu nowej cechy opisującej, tworzonej najczęściej przez (liniową) kombinację cech już istniejących. Można bowiem oczekiwać, że operacja ta spowoduje transformację zbiorów formalnie nieseparowalnych w zbiory liniowo-rozdzielne, lub prawie liniowo-rozdzielne. Ważną rolę w podwyższeniu efektywności badanego klasyfikatora, odgrywa właściwy dobór zbioru nauczonych wektorów wagowych, co zostało już stwierdzone w dotychczasowych badaniach. W najbliższym okresie czasu zostaną podjęte próby usprawnienia zdolności klasyfikacyjnych wspomnianego zbioru wygenerowanych wektorów wagowych, poddając je darwinowskim i nie-darwinowskim operacjom genetycznym [21].

BIBLIOGRAFIA

- [1] Hippe Z.S., Błajdo P.: *From the research on a new kNN pattern recognition method*. W: Burczyński T., Cholewa W., Moczulski W. (Eds.) *Methods of Artificial Intelligence*, Silesian University of Technology Edit. Office, Gliwice 2002, pp. 181-185.
- [2] Lauria E.J., Tayi G.K.: *Bayesian Data Mining and Knowledge Discovery*, In: Wang J. (Ed.) *Data Mining: Opportunities and Challenges*, Idea Group Publishing, Hershey (PA) 2003, pp. 260-277.
- [3] Hippe Z.S., Wrzesień M.: *Some problems of uncertainty of data after the transfer from multi-category to dichotomic problem space*. In: Burczyński T., Cholewa W., Moczulski W. (Eds.), *Methods of Artificial Intelligence*, Silesian University of Technology Edit. Office, Gliwice 2002, pp. 185-189.
- [4] Hippe Z.S., Knap M., Paja W.: *From Research on Pre-processing and Post-processing of Data in the Process of Creation Quasi-optimal Decision Trees*, In: Burczyński T., Cholewa W., Moczulski W. (Eds.) *Methods of Artificial Intelligence*, Silesian University of Technology Edit. Office, Gliwice 2002, pp. 177-180.
- [5] Pawlak Z.: *Wiedza a zbiory przybliżone*, W: Traczyk W. (Red.) *Problemy sztucznej inteligencji*, Wiedza i Życie, Warszawa 1995.
- [6] Varmuza K.: *Chemometrics: Multivariate View on Chemical Problems* In: Schleyer P. v. R., Allinger N. L., Clark T., Gasteiger J., Kollman P.A., Schafer III H. F., Schreiner P. R. (Eds.) *The Encyclopedia of Computational in Chemistry*, J. Wiley & Sons Ltd, Chichester 1998, Vol. 1, pp. 346-366.
- [7] Duda R.O., Hart P.E., Strok D.G., *Pattern Classification*, J. Wiley & Sons, New York 2001.
- [8] Hippe Z.S., Wrzesień M., *Z badań nad rozszerzeniem możliwości klasyfikacji liniowej maszyny uczącej*, W: Tadeusiewicz R., Ligęza L., Szymkat M., (Red.), *Metody i systemy komputerowe w badaniach naukowych i projektowaniu inżynierskim*, mat. III Krajowej Konferencji, Kraków 19-21 Listopad 2001.
- [9] Jurs P.C., Isenhour T.L., *Chemical Applications of Pattern Recognition*, J.Wiley & Sons, New York 1975.
- [10] Varmuza K., *Pattern Recognition in Chemistry*, Springer – Verlag, Heidelberg 1980.
- [11] Jelonek J., Stefanowski J., *Feature selection in the n^2 -classier applied for multiclass problems*, W: Burczyński T., Cholewa W., Moczulski W. (Eds.), *Methods of Artificial Intelligence*, PAS&SCSR, Gliwice 2002.
- [12] Jelonek J., *Rozprawa doktorska: Zastosowanie złożonego systemu klasyfikatora n^2 z mechanizmem konstruktywnej indukcji cech do wieloklasowych problemów uczenia maszynowego*, Politechnika Poznańska, Poznań 2000.
- [13] Grzymała-Busse J.W., Hippe Z.S., Wrzesień M.: *Klasyfikacja wielokategoryjnych przypadków przy użyciu klasyfikatora binarnego – opracowanie metodyki oraz narzędzia informatycznego*, [W:] Nycz M., Owoc M.L. (Red.), *Pozyskiwanie wiedzy i zarządzanie wiedzą*, Wyd. Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2003, s. 121-127.
- [14] <http://www.ics.uni.edu/~mlearn/>

- [15] Grzymała-Busse J.W., Hippe Z.S., *Data Mining Experiments with Melanoma Data Set*. In: *Intelligent Information Systems*, Kłopotek M., Michalewicz M., Wierchoń S.T. (Eds.), Physica-Verlag, Heidelberg 2000, pp. 27-34.
- [16] <http://www.stat.wisc.edu/~limt/>
- [17] Hippe Z.S., Iwaszek G., *From Research on a New Method of Development of Quasi – optimal Decision Trees*, w: Kłopotek M., Michalewicz M., Wierchoń S.T. (Red.), *Intelligent Information Systems IX*, Inst. Podstaw Informatyki PAN, Warszawa 2000, s. 31-35.
- [18] <http://www-ai.ijs.si/BlazZupan/car.html>
- [19] Hippe Z.S., Mroczek T., *Melanoma Classification and Prediction Using Belief Networks*, 3rd Conference on Computer Identification Systems KOSYR'2003, Miłków, 26-29.05.2003.
- [20] Kubat M., Bratko I., Michalski R.S., *A Review of Machine Learning Methods*, In: *Machine Learning and Data Mining: Methods and Applications*, Michalski R.S., Bratko I., Kubat M. (Eds.), J. Wiley & Sons, London 1998, pp. 3-69.
- [21] Michalski R.S., *Seminarium: Problemy generacji wiedzy: metodologia i bieżące badania*, Rzeszów 1.12.2003.

OPTIMIZATION OF SOME FUNCTIONS OF BINARY CLASSIFICATION IN PROCESS OF DATA MINING FROM MULTI-CATEGORY DATASETS

Summary

Selected features of binary classification machine used to data acquisition from multi-category data sets are shortly discussed. In the main part of paper the primary software modules are described. Then application of the first step of optimization learning process is presented. At the end of paper the results of experiments and conclusions are discussed.

Małgorzata Gajewska, Sławomir Gajewski

Katedra Systemów i Sieci Radiokomunikacyjnych, Politechnika Gdańska

GSM CORDLESS TELEPHONY SYSTEM – PERSPEKTYWA ROZWOJU TRANSMISJI DANYCH W SYSTEMIE GSM

Streszczenie

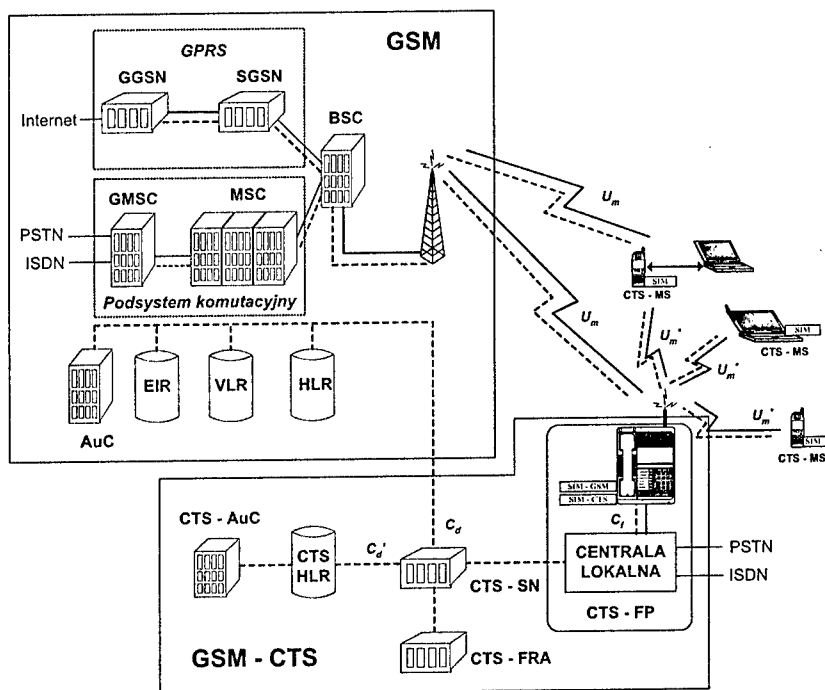
W artykule opisano nowy standard systemu telefonii bezprzewodowej, znanego pod nazwą GSM – CTS (ang. *GSM – Cordless Telephony System*), w pełni zintegrowanego z systemem GSM. Przedstawiono podstawowe charakterystyki omawianego systemu, tzn. jego architekturę, zasadę działania, właściwości interfejsu radiowego oraz rodzaje oferowanych usług. W referacie zostały omówione perspektywy rozwoju systemu, ze szczególnym uwzględnieniem możliwości jego zastosowania jako wyspecjalizowanego podsystemu systemu trzeciej generacji UMTS.

1. WSTĘP

Ewolucja systemów radiokomunikacyjnych drugiej generacji zmierzająca do budowy Uniwersalnego Systemu Radiokomunikacji Ruchomej trzeciej generacji UMTS oznacza także daleko posuniętą integrację systemów radiokomunikacyjnych różnego przeznaczenia, a w tym systemów telekomunikacji bezprzewodowej, komórkowej oraz w przyszłości także satelitarnej. Rozwój systemu GSM, który obecnie jest realizowany w ramach tzw. fazy 2+, uwidacznia w pełni te trendy, a jednym z jego przejawów jest powstający standard systemu telefonii bezprzewodowej nazwanego systemem GSM – CTS. Jest to nowoczesny system przeznaczony do transmisji sygnałów mowy i danych, w którym będą dostarczane usługi telefonii bezprzewodowej do sieci stałej oraz sieci komórkowej. W dotychczasowych systemach komórkowych były wyraźnie rozdzielone zastosowania publicznych, rozległych oraz lokalnych systemów, skupionych na małych obszarach (np. systemu DECT). System GSM – CTS ma zapewnić integrację funkcjonalności rozległych systemów telefonii komórkowej oraz lokalnych systemów telefonii bezprzewodowej, przy jednoczesnym zapewnieniu możliwości korzystania z obu systemów przy użyciu tego samego terminala. Dzięki temu nastąpi znaczące rozszerzenie oferty operatorów, przede wszystkim w zakresie szybkiej transmisji danych. Jednocześnie użytkowanie terminali stanie się bardziej uniwersalne. Przystosowanie terminala wykorzystywanego w lokalnej łączności bezprzewodowej do standardu stacji ruchomej systemu GSM będzie odbywać się przy niewielkiej ingerencji w strukturę tej stacji, ograniczonej do rozszerzenia funkcjonalności oprogramowania oraz zawartości karty SIM. Będzie to możliwe wyłącznie w przypadku, gdy budowa interfejsu radiowego systemu GSM – CTS będzie bardzo zbliżona do interfejsu radiowego systemu GSM.

2. ARCHITEKTURA SYSTEMU GSM – CTS

Bezprzewodowy system komórkowy GSM – CTS jest systemem bazującym na interfejsie radiowym systemu GSM, który umożliwia komunikację pomiędzy lokalnymi radiowymi podsystemami stałymi (ang. *CTS – Fixed Part*), a stacjami ruchomymi CTS – MS (ang. *CTS – Mobile Station*) [1, 2]. Radiowe podsystemy stałe tworzą lokalne centrale (ang. *Local Exchange*) oraz urządzenia nadawczo-odbiorcze lokalnych stacji bazowych. Podsystem stały ma z reguły połączenia stałe z terminalami telefonicznymi, przewodowymi lub bezprzewodowymi.



Rys.1. Ogólna architektura systemu GSM – CTS na tle architektury systemu GSM

Każdy podsystem stały CTS – FP jest połączony łączem stałym z publiczną siecią telefoniczną PSTN (ang. *Public Switched Telephone Network*) lub siecią ISDN (ang. *Integrated Service Digital Network*) lub łączem radiowym z siecią GSM. Ponadto przy pewnym wyposażeniu uzupełniającym jest możliwa komunikacja bezpośrednio z sieciami z komutacją pakietów, np. Internetem. W przypadku, kiedy stacja bazowa podsystemu CTS – FP realizuje połączenie z siecią stałą (PSTN lub ISDN), to wówczas jej działanie jest oparte na komutacji kanałów i sygnalizacji charakterystycznej dla sieci stałej, przy czym nie ma bezpośredniej komunikacji radiowej pomiędzy różnymi stacjami bazowymi. Natomiast, jeśli stacja bazowa podsystemu CTS – FP realizuje połączenie z siecią komórkową, to wówczas w grę wchodzi sterowanie transmisją charakterystyczne dla systemu komórkowego GSM. Z punktu widzenia sieci GSM stacje bazowe systemu CTS są wówczas widziane jako standardowe stacje ruchome systemu GSM, zawierające typowe dla tego systemu karty SIM. Stacje bazowe podsystemu CTS – FP mogą niekiedy zawierać

kartę SIM charakterystyczną dla systemu CTS [1]. Ogólną architekturę systemu GSM – CTS przedstawiono na rys. 1.

Podstawowe elementy architektury systemu CTS stanowią węzły CTS – SN (ang. *Service Node*) oraz CTS – FRA (ang. *CTS Frequency Allocation*). Pierwszy z tych węzłów jest węzłem usługowym i pełni funkcje połączeniowe, natomiast drugi jest wykorzystywany w procesie alokacji częstotliwości. Węzeł CTS – SN łączy bazę danych abonentów macierzystych CTS – HLR (ang. *CTS Home Location Register*) ze stacjami bazowymi podsystemów CTS – FP poprzez sieć dostępową. Węzeł ten pełni również funkcje ochrony informacji sygnalizacyjnych dotyczących interfejsu stałego, który stanowi połączenie pomiędzy stacją bazową, a siecią stałą i komórkową, umożliwiające komunikację pomiędzy użytkownikami. W wymianie informacji pomiędzy węzłem CTS – SN i stacją bazową podsystemu CTS – FP oraz pomiędzy węzłem CTS – SN i stacją ruchomą CTS – MS, pośredniczy lokalna centrala. Natomiast węzeł CTS – FRA jest wykorzystywany w procesie przydziału częstotliwości przy różnych operacjach realizowanych przez system CTS [2].

Baza danych użytkowników macierzystych CTS – HLR zawiera różnorakie dane dotyczące abonentów zarejestrowanych w obszarze działania stacji bazowej obsługiwanej przez danego operatora systemu CTS, którym może być operator systemu GSM lub jakikolwiek inny. Baza danych CTS – HLR służy do zarządzania danymi abonentów z uwzględnieniem ich odpowiedniej ochrony oraz do przechowywania dodatkowych informacji dotyczących poszczególnych abonentów. Bezpośrednio z nią współpracuje centrum uwierzytelniania AuC systemu CTS (ang. *CTS Authorization Center*), które realizuje odpowiednie przetwarzanie danych związane z autoryzacją karty SIM stacji bazowej CTS – FP [2].

Bardzo istotną cechą systemu GSM – CTS jest możliwość bezpośredniego wykorzystywania infrastruktury sieciowej systemu GSM, a w szczególności baz danych użytkowników macierzystych HLR, wizytujących VLR (ang. *Visitors Location Register*) i bazy danych sprzętu radiokomunikacyjnego EIR (ang. *Equipment Identity Register*) oraz centrum uwierzytelniania AuC. Dzięki temu nie ma konieczności budowania lokalnej infrastruktury systemu CTS przeznaczonej do realizacji funkcji tych elementów sieci.

Ze względu na integrację systemu GSM – CTS z systemem GSM stacje ruchome MS – CTS będą terminalami, z których korzystają abonenci systemu GSM, lecz odpowiednio przystosowanymi do dodatkowych zadań. Każda stacja ruchoma MS – CTS będzie się składać z terminala systemu GSM oraz z karty GSM SIM, przy czym oba te urządzenia będą programowo adoptowane dla potrzeb systemu CTS. Zadaniem każdej stacji ruchomej będzie umożliwienie użytkownikom bezprzewodowego dostępu do sieci CTS oraz sieci komórkowej GSM.

Na rys. 1 zaznaczono dwa typy interfejsów. Liniami przerywanymi oznaczono interfejsy wykorzystywane do przesyłania informacji sygnalizacyjnych pomiędzy różnymi elementami architektury systemu CTS, natomiast linią ciągłą oznaczono interfejs wykorzystywany do komunikacji pomiędzy poszczególnymi użytkownikami systemu. Interfejs radiowy U_m systemu CTS jest wykorzystywany do lokalnej komunikacji pomiędzy stacjami ruchomymi CTS – MS, a stacjami bazowymi podsystemu CTS – FP. Realizuje on bezprzewodowe połączenie pomiędzy CTS – MS a CTS – FP, a w przyszłości ma zostać wykorzystany do realizacji usług przesyłania danych. Interfejs radiowy U_m jest, zmodyfikowanym w niewielkim stopniu, interfejsem radiowym systemu GSM. Interfejs C_d pomiędzy węzłem usługowym CTS – SN, a bazą danych CTS – HLR służy do przesyłania informacji uwierzytelniających zawartych na karcie CTS SIM stacji bazowych. Natomiast interfejs C_d pomiędzy węzłem CTS – SN, a bazą danych abonentów macierzystych systemu

GSM umożliwia autoryzację danych zawartych w karcie GSM SIM w przypadku, kiedy tego typu weryfikacja jest realizowana z wykorzystaniem sieci CTS. Interfejs C_f łączy węzeł CTS – SN ze stacją bazową podsystemu stałego CTS – FP [2].

Możliwa jest modyfikacja architektury systemu CTS polegająca na tym, że stacja bazowa CTS – FP może realizować bezpośrednie połączenia z siecią GSM. Wówczas interfejs C_f pomiędzy CTS – FP, a siecią stałą (PSTN) zostaje zamieniony na standardowy interfejs radiowy U_m systemu GSM, a stacja bazowa systemu CTS jest widziana, z punktu widzenia sieci GSM, jako stacja ruchoma systemu GSM, czyli zwyczajny terminal systemu GSM z odpowiednią kartą GSM SIM [2].

Ponadto możliwa jest realizacja hierarchicznej architektury systemu CTS, w której będzie możliwe wykorzystanie tzw. podsystemu nadzorującego oraz kilku podsystemów lokalnych. W skład każdego podsystemu lokalnego będą wchodziły podsystemy stałe CTS – FP, wraz ze stacjami bazowymi oraz przyporządkowane do tych podsystemów stacje ruchome. Komunikacja pomiędzy tymi stacjami będzie odbywała się poprzez interfejs radiowy. Natomiast podsystem nadzorujący będzie pełnił funkcję scentralizowanego systemu nadzoru, w którym będzie realizowany przydział częstotliwości dla odpowiednich podsystemów stałych CTS – FP, poprzez węzeł CTS – FRA oraz operacje związane z uwierzytelnianiem abonentów i aktualizacją baz danych [2].

3. INTERFEJS RADIOWY SYSTEMU GSM – CTS

Interfejs radiowy systemu CTS, czyli interfejs radiowy pomiędzy stacją ruchomą CTS – MS, a podsystemem stałym CTS – FP, będzie zmodyfikowanym w niewielkim stopniu interfejsem radiowym systemu GSM. Oznacza to, że podobnie jak w systemie GSM zastosowana zostanie technika wspólnego użytkowania kanału częstotliwościowego TDMA (ang. *Time Division Multiple Access*), polegająca na niezależnym przesyłaniu sygnałów wielu użytkowników w odpowiednich niezależnych ramach czasowych. Ponadto transmisja sygnałów odbywać się będzie w trybie duplexu częstotliwościowego FDD (ang. *Frequency Division Duplex*), tzn. w dwóch niezależnych podpasmach częstotliwości przeznaczonych osobno do nadawania i odbioru. System CTS będzie wykorzystywał częstotliwości, które zostały przydzielone systemowi GSM pracującemu w paśmie 900 MHz oraz 1800 MHz z odległością międzykanałową 200 kHz [3]. Dodatkowo, oprócz wykorzystywania częstotliwości oferowanych przez system GSM, system CTS będzie mógł operować na częstotliwościach, które zostaną mu przydzielone poprzez indywidualną licencję. W systemie GSM – CTS będzie stosowany ten sam typ modulacji, tzn. GMSK lub 8PSK oraz identyczne schematy kodowania kanałowego i przeplotu, jak w systemie GSM [4], wraz z ich ewentualnymi modyfikacjami realizowanymi w ramach budowy podsystemu EDGE (ang. *Enhanced Data Rates for Global Evolution*).

4. KOMPATYBILNOŚĆ ELEKTROMAGNETYCZNA

Interfejs radiowy systemu CTS został zaprojektowany w taki sposób, aby zapewnić możliwie małe poziomy interferencji wytwarzanych, zarówno w pasmach wspólnych systemów CTS i GSM, jak i interferencji wchodzących w grę pomiędzy różnymi systemami CTS. Taki rezultat można uzyskać przez stosowanie kierunkowych wiązek promieniowania anten (ang. *Beacon Concept*), adaptacyjnej alokacji częstotliwości AFA (ang. *Adaptive Frequency Allocation*) oraz hoppingu częstotliwościowego TFH (ang. *Total Frequency Hopping*) [4].

Do każdego podsystemu stałego będzie przyporządkowana ograniczona liczba stacji ruchomych, dlatego w systemie CTS nie będzie konieczności korzystania z rozświeczanego kanału sygnalizacyjnego BCCH (ang. *Broadcast Control Channel*) zdefiniowanego dla systemu GSM, który służy do przekazywania informacji systemowych. W systemie CTS zaproponowano zastosowanie kanału o transmisji ukierunkowanej przestrzennie CTS-BCH (ang. *CTS Beacon Channel*), w którym są transmitowane sygnały pozwalające na przydział właściwej częstotliwości dla każdej stacji ruchomej oraz na uzyskanie odpowiedniej synchronizacji. Na podstawie tych informacji stacja ruchoma CTS – MS jest w stanie dokonać synchronizacji oraz zidentyfikować stację bazową podsystemu CTS – FP. Mechanizm ten został zastosowany w celu zmniejszenia mocy sygnałów, emitowanych zarówno przez stację bazową, jak i stację ruchomą i dlatego jest to jedyny kanał logiczny, którego sygnały są wysyłane regularnie – sygnały z pozostałych kanałów logicznych są przesyłane „na żądanie” [4].

Zaproponowano także zastosowanie w stacji bazowej mechanizmu adaptacyjnej alokacji częstotliwości AFA, które ma na celu ograniczenie ryzyka wystąpienia interferencji wspólnego- i sąsiedniokanałowych, pomiędzy sygnałami przesyłanymi w sieci CTS i sieci GSM. W ogólności AFA polega na wybieraniu przez stację bazową CTS – FP odpowiednich częstotliwości zapisanych na tzw. głównej liście częstotliwości GFL (ang. *Generic Frequency List*) w taki sposób, aby pomiędzy transmisją w łączu w górę (czyli od stacji ruchomych do stacji bazowej) i w łączu w dół (czyli od stacji bazowej do stacji ruchomych) występowały możliwie małe interferencje [4]. Główna lista częstotliwości GFL zawiera numery wszystkich kanałów częstotliwościowych (ang. ARFCN – *Absolute Radio Frequency Channel Number*), z których mogą korzystać operatorzy stacji bazowych podsystemów stałych systemu CTS. W wyniku pomiarów poziomu mocy interferencji na poszczególnych częstotliwościach z listy GFL zostanie stworzona tabela przydziału częstotliwości, w której w określonym porządku będą zapisane częstotliwości dostępne dla systemu. Z listy tej będzie wynikać, które częstotliwości nie nadają się do zastosowania w systemie CTS, z uwagi na zbyt duże moce sygnałów zakłócających (zwłaszcza pochodzących od systemu GSM).

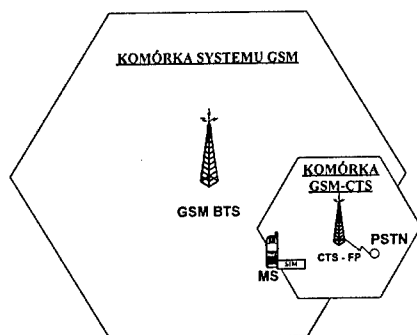
Ograniczenie interferencji może odbywać się dzięki wykorzystaniu mechanizmu hoppingu częstotliwościowego TFH, zwłaszcza pomiędzy systemem CTS, a występującymi na danym obszarze nakładkowymi systemami komórkowymi [4]. Zastosowanie mechanizmów skokowej zmiany częstotliwości nośnych podczas transmisji sygnałów spowoduje zmniejszenie wzajemnego wpływu sygnałów interferujących, których moc jest w ten sposób wtrącana do różnych połączeń oraz ograniczenie wpływu szybkich i powolnych fluktuacji sygnału na jakość transmisji, tj. wpływu zaników i zmieniającej się z odległością mocy średniej sygnału.

5. USŁUGI W SYSTEMIE GSM – CTS

Rodzaje usług, które oferuje system GSM – CTS, są zależne od tego, z jakiego typu sieciami są połączone stacje bazowe podsystemów CTS – FP. Aby oferować określone usługi, operator systemu CTS musi mieć zgodę operatora systemu GSM na użytkowanie przez stację bazową częstotliwości, które są wykorzystywane w sieci komórkowej. Operatorzy systemu CTS w porozumieniu z operatorami systemu GSM muszą określić granice obszarów, na których będą oferować swoje usługi. Oczywiście operator systemu CTS może być równocześnie operatorem systemu GSM [2]. W praktyce można wyróżnić trzy różne przypadki:

- zasięg systemu CTS jest całkowicie pokryty zasięgiem sieci komórkowej,
- zasięg systemu CTS jest tylko w pewnym stopniu pokryty zasięgiem sieci komórkowej,
- obszar działania systemu CTS jest całkowicie poza zasięgiem działania systemu GSM.

Przykładowo, koncepcję działania systemu CTS, znajdującego się w obszarze działania systemu GSM, w przypadku połączenia systemu CTS bezpośrednio z publiczną siecią telefoniczną PSTN, przedstawiono na rys. 2. Każda stacja ruchoma CTS – MS, która znajdzie się w zasięgu stacji bazowej systemu CTS, może komunikować się z nią z pominięciem sieci komórkowej GSM. Wówczas użytkownik może przyjąć i zaakceptować wywołanie pochodzące bezpośrednio od sieci PSTN, bez udziału sieci komórkowej. Możliwa jest także jednoczesna komunikacja stacji ruchomej ze stacją bazową systemu CTS oraz stacją bazową systemu GSM, realizowana w tzw. strukturze parasolowej. Jeżeli natomiast stacja ruchoma znajdzie się poza zasięgiem stacji bazowej systemu CTS, to przyłączana jest automatycznie do sieci komórkowej GSM. Odpowiednia sygnalizacja informuje użytkownika, do jakiej sieci jest aktualnie przyłączony [1].

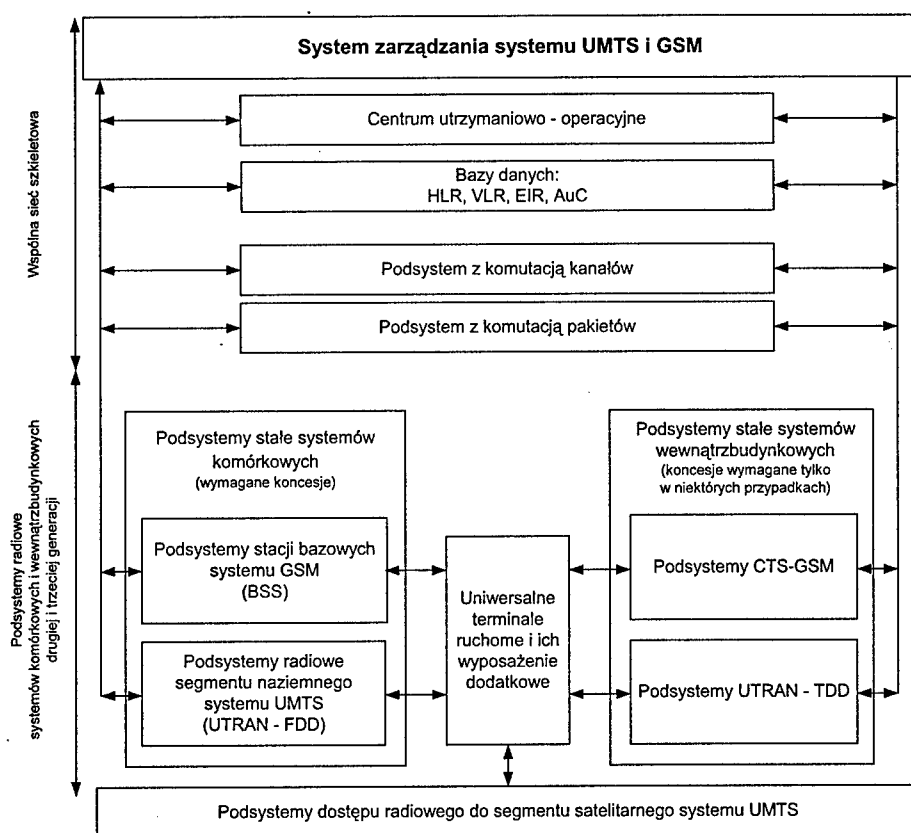


Rys.2. Koncepcja współdziałania systemu GSM i CTS [1]

W ogólności, w systemie CTS będą dostępne te same usługi, jak w systemie GSM, a więc zarówno transmisja sygnałów mowy, jak i danych, w trybie komutacji kanałów i pakietów [1], włącznie z najnowszymi rozwiązaniami stosowanymi w podsystemie EDGE. Dodatkowo, dzięki umożliwieniu bezpośredniego przyłączenia podsystemów stałych do sieci przewodowych, usługi te będą mogły być wzbogacone o usługi szerokopasmowe, które nie będą oferowane bądź będą udostępnione w ograniczonym stopniu, w sieci GSM, np. bezpośredni dostęp do Internetu. Będzie to jednak wymagało rozbudowy funkcji podsystemów stałych CTS – FP. Jest to bardzo ważna cecha systemu CTS, ponieważ oznacza możliwość przesyłania danych drogą radiową z pominięciem terminala ruchomego. Może być ona realizowana z wykorzystaniem stacji bazowych podsystemu CTS – FP. Transmisja ta, dzięki małym odległościom pomiędzy np. laptopem i stacją bazową systemu CTS, mogłaby odbywać się ze znacznymi szybkościami transmisji, nieosiągalnymi przy komunikacji ze stacjami bazowymi systemu GSM. Usługi dodatkowe żądane przez stacje ruchome będą mogły być realizowane za pośrednictwem stacji bazowej systemu CTS lub stacji bazowej systemu GSM, w zależności od potrzeb i możliwości technicznych na danym obszarze.

6. FUNKCJE SYSTEMU GSM – CTS W SYSTEMIE UMTS

W obliczu zachodzącego obecnie procesu integracji systemów radiokomunikacyjnych i ich ewolucji zmierzającej do powstania jednego globalnego systemu UMTS, system telekomunikacji bezprzewodowej CTS, podobnie jak inne systemy, będzie posiadał swoje miejsce w globalnej architekturze systemu trzeciej generacji, co pokazano na rys. 3.



Rys.3. Ogólna architektura systemu trzeciej generacji z uwzględnieniem systemów radiokomunikacji ruchomej i systemów telekomunikacji bezprzewodowej.

Całkowity proces przemian podsystemów składowych systemu GSM, a przede wszystkim niedawno powstałego podsystemu GPRS do transmisji danych w trybie komutacji pakietów oraz obecnie wprowadzanego podsystemu EDGE, zmierza w kierunku udostępnienia usług przesyłania danych ze znacznymi szybkościami transmisji, sięgającymi od kilkudziesięciu do kilkuset kb/s. Jednak spełnienie tych wymagań będzie znacząco utrudnione na obszarach otwartych, natomiast jest zupełnie realne w systemach wewnętrznych, gdzie jest możliwe uzyskanie wysokiej jakości transmisji, głównie ze względu na stosunkowo krótkie odległości pomiędzy stacjami ruchomymi a bazowymi. System CTS może zatem wypełnić lukę panującą w infrastrukturze systemu GSM i usprawnić komunikację w ramach systemu bezprzewodowego, wypierając niektóre obecnie coraz bardziej

popularne rozwiązania, jak np. system DECT, dzięki swojej elastyczności i integralności z systemem GSM, którego zasięg działania jest w Europie powszechny. System CTS nie zastąpi rozwiązań, na które pozwoli w przyszłości interfejs radiowy systemu UMTS, zarówno stosowany globalnie standard UTRA – FDD (ang. *UMTS Terrestrial Radio Access*) w trybie duplexu częstotliwościowego, jak również, przeznaczony do pracy wewnątrz budynków, standard UTRA – TDD, pracujący w trybie duplexu czasowego. Jednak ze względu na znaczne koszty wprowadzanie podsystemów wewnątrzbudynkowych systemu UMTS będzie znacząco opóźnione, a tanie i proste rozwiązania systemu CTS, mogą bardzo szybko znaleźć swoich użytkowników.

7. ZAKOŃCZENIE

Zgodnie z założeniami projektu GERAN (ang. *GSM / EDGE Radio Access Network*), realizowanego w grupie standaryzacyjnej 3GPP (*3rd Generation Partnership Project*) systemu UMTS, system CTS może znaleźć swoje miejsce w globalnej architekturze systemu UMTS. Może on do czasu upowszechnienia się systemu UMTS, a w szczególności jego podsystemów przeznaczonych do pracy wewnątrz budynków, stać się standardem telefonii bezprzewodowej, decydującym w istotny sposób o nowej jakości usług przesyłania sygnałów mowy, obrazu oraz danych drogą radiową. Szanse systemu CTS są tym większe, że koszty jego wdrażania będą bardzo niskie, a wykorzystywanie podsystemu bardzo wygodne dla firm i użytkowników indywidualnych, ze względu na możliwość korzystania z tych samych terminali, które są stosowane powszechnie w systemie GSM. Ponadto istnieje możliwość jego rozpowszechnienia wśród operatorów prywatnych, niezwiązanych z systemem GSM (pod warunkiem przydzielenia nowego pasma częstotliwości) oraz możliwość zawierania umów z operatorami GSM. Możliwe jest także rozszerzenie oferty operatorów GSM o nowe podsystemy i usługi oraz obniżenie kosztów eksploatacyjnych, a ponadto usprawnienie komunikacji z pominięciem kablowych sieci telefonicznych PSTN oraz sieci ISDN.

BIBLIOGRAFIA

- [1] 3GPP TSG SSA Technical Specification 42.056 v.5.0.0, GSM Cordless Telephony System (CTS), Phase 1, Service description. Stage 1. Release 5, 06-2002.
- [2] ETSI EN European Standard 302 405 v7.1.1, GSM Cordless Telephony System (CTS), Phase 1, CTS Architecture Description. Stage 2, GSM 03.56, Release 1998, 08-2000.
- [3] 3GPP TSG GERAN Technical Specification 45.056 v.5.0.0, GSM Cordless Telephony System (CTS), Phase 1, CTS-FP Radio subsystem. Release 5, 06-2002
- [4] 3GPP TSG GSM/EDGE RAN Technical Specification 43.052 v.5.0.0, GSM Cordless Telephony System (CTS), Phase 1, Lower Layers of the CTS Radio Interface. Stage 2, Release 5, 06-2002.

GSM-CTS – A PERSPECTIVE OF DATA TRANSMISSION EVOLUTION IN GSM

Summary

In the paper a new standard of wireless communication system, integrated with mobile GSM network, called GSM – CTS (GSM Cordless Telephony System) has been presented. Basic characteristics of GSM – CTS including an architecture, functional description, radio interface and offered services have been described. The perspectives of its introduction as a specialized subsystem and the integration with third generation UMTS system have been outlined.

Janusz Jurski, Józef Woźniak

Katedra Systemów Informacyjnych, Politechnika Gdańska

BADANIA ALGORYTMÓW WYBORU TRAS W NISKOORBITOWYCH SIECIACH SATELITARNYCH

Streszczenie

Artykuł przedstawia wyniki badań symulacyjnych różnych algorytmów wyboru tras (ang. *routing*) w niskoorbitowych sieciach satelitarnych. Badania uwzględniają nierównomierny w skali globu rozkład zapotrzebowania na usługi transmisji danych. Właściwości wybranych algorytmów badano przy różnym stopniu obciążenia sieci satelitarnej. Obserwowano rozkład obciążenia sieci, opóźnienia transmisji oraz ilość pakietów traconych wskutek przepełniania się kolejek.

1. WSTĘP

Niskoorbitowe sieci satelitarne (ang. Low-Earth Orbit – LEO) posiadają zazwyczaj gęstą sieć bezpośrednich połączeń międzysatelitarnych (ang. Inter-Satellite Link – ISL). Połączenia międzysatelitarne umożliwiają kierowanie ruchu trasą o długości liczonej jako odległość fizyczna dużo mniejszej niż jest to możliwe w innych rodzajach sieci (inne rodzaje sieci satelitarnych, a nawet sieci naziemne). Oznacza to mniejsze opóźnienia propagacji, a więc potencjalnie mniejsze opóźnienia transmisji.

Gęsta sieć połączeń między węzłami sieci stwarza także korzystne warunki dla mechanizmów kierowania ruchem. W takiej sieci istnieje bowiem zazwyczaj kilka tras bardzo zbliżonych do trasy optymalnej pod względem opóźnienia. Ta cecha odróżnia sieci satelitarne od sieci naziemnych, gdzie zazwyczaj (w skali sieci globalnej) istnieje jedna trasa optymalna a inne wprowadzają opóźnienia dużo większe.

Z drugiej strony, ze względów ekonomicznych, jak również z powodu zwiększającego się stale zapotrzebowania na ilość transportowanych danych, należy spodziewać się, że sieci satelitarne będą wykorzystywane przy dużym obciążeniu – wykorzystując w możliwie największym stopniu przepustowość połączeń ISL. Zważywszy jednocześnie, że ruch w skali globu transmitowany jest zazwyczaj pomiędzy dużymi skupiskami ludzi o dużym stopniu rozwoju cywilizacyjnego, takimi jak Europa, Japonia, czy pewne regiony Stanów Zjednoczonych, należy spodziewać się dużych nierównomierności w obciążeniu sieci, a nawet przeciążeń prowadzących do przepełnienia kolejek i utraty pakietów.

Mając powyższe na uwadze, przeprowadzono badania symulacyjne dla niskoorbitowej sieci satelitarnej wzorowanej na znanej sieci Iridium (z niewielkimi modyfikacjami, opis-

anymi w rozdziale 2). Wykorzystano w tym celu ogólnodostępny i często używany w środowisku naukowym symulator o nazwie NS (skrót od ang. *Network Simulator*). Zasymulowano działanie kilku różnych algorytmów wyboru tras, opisanych w rozdziale 4. Ich działanie badane było przy zmieniającym się obciążeniu sieci satelitarnej. Obciążenie generowano w sposób możliwie przypominający rzeczywisty rozkład ruchu na Ziemi, jak opisuje to rozdział 3. Wyniki badań i ich analiza znajdują się w rozdziałach 5 i 6.

2. SYMULOWANA SIEĆ SATELITARNA

Sieć satelitarna, dla której przeprowadzano badania symulacyjne, była wzorowana na sieci Iridium, której budowę w 1986 r. zainicjowała firma Motorola. Satelity tej sieci krążą na wysokości 780 km nad powierzchnią Ziemi, na sześciu orbitach. W sumie w sieci jest 66 satelitów – po 11 na każdej orbicie. Można dodać, że okres obiegu satelity wynosi 6026,9 sekund (1 godz. 40 minut 26,9 sekund).

Satelity w ramach jednej orbity oddalone są od siebie o $32,7^\circ$. Natomiast odległość pomiędzy sąsiednimi orbitami wynosi $31,6^\circ$. Orbits nie przebiegają dokładnie nad biegunem Ziemi – ich kąt inklinacji¹ wynosi $86,4^\circ$.

Przy takich parametrach orbit jak powyższe, sieć satelitarna podzielona jest na dwie płaszczyzny, w których satelity krążą po orbitach w tym samym kierunku. Natomiast na styku tych płaszczyzn – nazywanym dalej szwem (ang. *seam*) – satelity poruszają się względem siebie w przeciwnych kierunkach. W sieci Iridium odległość kątowa pomiędzy orbitami znajdującymi się po przeciwnych stronach szwu wynosi 22° .

W symulowanej sieci satelity posiadają cztery łącza do sąsiednich satelitów: dwa łącza do satelity poprzedzającego i następnego na tej samej orbicie (ang. *intra-plane ISL*) oraz dwa łącza do satelitów na sąsiednich orbitach (ang. *inter-plane ISL*). W prawdziwej sieci Iridium, satelity znajdujące się przy szwie posiadają tylko trzy łącza, gdyż łącza poprzez szew (ang. *cross-seam ISL*) nie są utrzymywane. Ponieważ jednak w nowszych, planowanych sieciach satelitarnych (np. sieć *Teledesic*) łącza poprzez szew są przewidziane, to także w sieci symulowanej je wprowadzono. Łącza przez szew w znaczny sposób poprawiają łączność pomiędzy płaszczyznami po dwóch stronach szwu – jeżeli ich nie ma, to obydwie płaszczyzny sieci są ze sobą połączone tylko w okolicy biegunów. Z drugiej strony, łącza przez szew wymagają częstego przełączania (ang. *handoff*), co odróżnia je od łączy wewnątrz płaszczyzny, które co najwyżej muszą być czasowo wyłączane w okolicach biegunów Ziemi, ale nie muszą być nigdy przełączane do innych satelitów.

Łącza Ziemia-satelita (ang. *Ground-Satellite Link – GSL*) również wymagają przełączania. W sieci Iridium założono minimalny kąt elewacji² równy $8,2^\circ$. W symulacjach oznaczało to, że jeżeli satelita znajdował się poniżej tej wysokości, to wyszukiwany był inny satelita, z którym łączność była możliwa i zestawiano z nim łącze. W prawdziwej sieci Iridium do łączności z obiektami naziemnymi stosowana jest zasada wielodostępu z podziałem częstotliwości i wielodostępu z podziałem czasowym FDMA/TDMA. Ponieważ badania miały skupić się na własnościach części satelitarnej i pomijać elementy związane z łącznością Ziemia-satelita, w symulowanej sieci mechanizmy te nie były zaimplementowane.

¹ Kąt inklinacji określa pochylenie orbity (tzw. orbity polarne, które przebiegają nad biegunem mają kąt inklinacji równy 90°).

² Kąt elewacji określa wysokość satelity nad horyzontem.

Nie były również symulowane niedokładności sieci satelitarnej, w szczególności niekołowe i niewspółśrodkowe orbity oraz odchylenia satelitów od ich pozycji nominalnych. Elementy te nie powinny mieć wpływu na ostateczne wyniki symulacji.

Natomiast istotnym parametrem mającym wpływ na własności sieci jest przepustowość (ang. bandwidth) łączy ISL i GSL. Prawdziwa sieć Iridium służy do przenoszenia ruchu telefonii satelitarnej z prędkościami będącymi wielokrotnością 2,4 kb/s.

W badaniach symulacyjnych starano się skupić się na własnościach algorytmów kierujących ruchem w sieci satelitarnej w sytuacji, gdy łączy ISL są mocno obciążone. Z powodu złożoności i czasochłonności badań symulacyjnych starano się zminimalizować liczbę pakietów niezbędną do obciążenia sieci, ale jednocześnie zadbane aby nie stracić możliwości porównania badanych algorytmów. Z tego względu przyjęto, że łączy ISL mają nierealnie małą przepustowość 0,5 kb/s. Pozwoliło to łatwo (małą ilością pakietów) doprowadzić sieć do obciążenia, przy którym przepełniały się kolejki i gubione były pakiety. Z powodu tych zmian uzyskane wartości opóźnień pakietów nie odpowiadają wartościom rzeczywistym, jednak ich względne zależności pozostają takie same, co jest istotne przy porównywaniu algorytmów wyboru tras. Dodatkowo starano się zminimalizować wpływ mechanizmów łączności Ziemia-satelita na wyniki badań – z tego względu przyjęto, że łączy GSL mają bardzo dużą przepustowość 1,5 Gb/s (choć to dodatkowo spowodowało, że uzyskane wartości opóźnień nie odpowiadają wartościom rzeczywistym).

Kolejki w symulowanej sieci miały długość maksymalnie 5 pakietów dla każdego łączy a pakiety były usuwane zgodnie z polityką tail-drop (czyli usuwano pakiety przychodzące, gdy kolejka miała już maksymalną długość).

3. MODEL RUCHU

Starano się zasymulować jak najbardziej rzeczywisty rozkład obciążenia sieci satelitarnej. W tym celu na całym globie rozmieszczono terminale oddalone od siebie o 15° długości i szerokości geograficznej. Terminale te wysyłały do siebie pakiety o długości 1000 bajtów, z odstępem opisanym rozkładem Poissona.

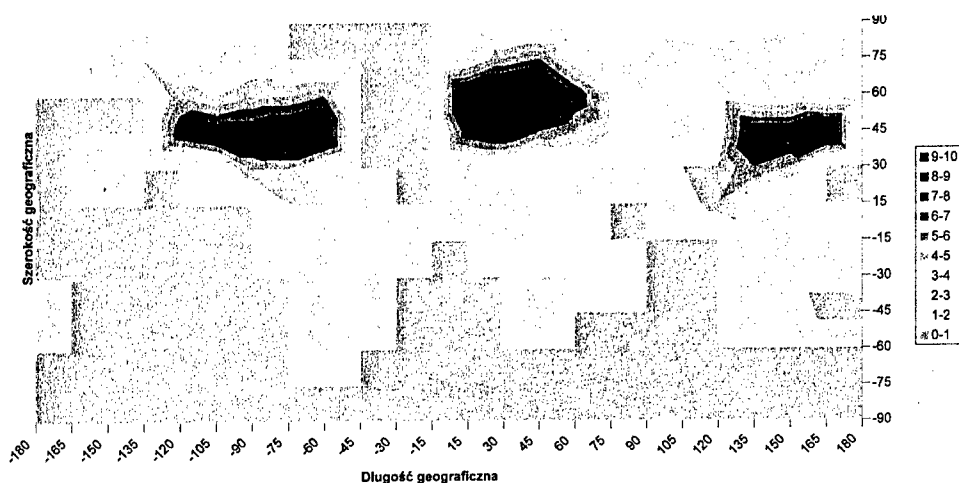
Dla każdej pary terminali starano się dobrać ilość generowanego ruchu tak, aby oddać rozkład ruchu na Ziemi. Posłużono się w tym celu modelem opisanym między innymi w [1] i [2], gdzie ilość ruchu pomiędzy dwoma węzłami sieci jest zależna od wartości potencjalnego zapotrzebowania w miejscu na Ziemi, w którym znajdują się te węzły oraz od odległości pomiędzy tymi węzłami.

Ilość ruchu pomiędzy węzłami A i B jest określona wzorem:

$$T(A, B) = \frac{(w_A \cdot w_B)^{0,6}}{(d(A, B))^{1,7}} \quad (3.1)$$

gdzie: w_A, w_B – potencjalne zapotrzebowania dla węzłów A oraz B ,
 $d(A, B)$ – odległość pomiędzy węzłami.

Zgodnie z [1], potencjalne zapotrzebowanie można opisać mapką, taką jak na rys. 1:



Rys.1. Potencjalne zapotrzebowanie na transmisję danych.

Autorzy w [1] i [2] pokazują, że zastosowany model ruchu dosyć dobrze odzwierciedla ruch generowany w rzeczywistości. W badaniach zmieniano sumaryczną ilość ruchu, nie zmieniając względnych proporcji w transmisji pomiędzy odpowiednimi regionami. Pozwoliło to porównać właściwości algorytmów przy różnym obciążeniu sieci.

4. BADANE ALGORYTMY WYBORU TRAS

Badania dotyczyły algorytmów wyboru tras opisanych w poniższych podrozdziałach. Przekazując pakiet każdy satelita stosował wybrany algorytm do wyznaczenia następnego satelity, do którego należało wysłać pakiet.

4.1. Min-delay

W algorytmie *min-delay* ruch kierowano trasą o najmniejszym czasie propagacji. Do wyliczenia najkrótszej drogi w grafie wykorzystano ogólnie znany algorytm Dijkstry. Czas propagacji sygnału był liczony na podstawie fizycznej odległości pomiędzy węzłami sieci.

4.2. Min-hop

W algorytmie *min-hop* ruch kierowany był trasą o najmniejszej liczbie węzłów na ścieżce – jest to algorytm dobrze znany z protokołów takich jak RIP.

4.3. Geograficzny

Algorytm routingu geograficznego (oznaczany dalej jako *geo*) został opisany w [3]. W algorytmie tym pakiet jest kierowany do tego węzła sąsiedniego, z którego jest najbliższy (w sensie odległości geograficznej liczonej w linii prostej) do celu.

Taki algorytm charakteryzuje się bardzo niewielką złożonością obliczeniową, jednak w niewielkiej odległości od celu może mieć problem ze znalezieniem następnego węzła, do

którego należy przekazać pakiet. Stąd też w [3] opisano nieco zmodyfikowaną wersję tego algorytmu nazywaną z ang. „locally scoped shortest path”. W wersji tej, w niewielkiej odległości od celu, stosowany jest po prostu algorytm *min-delay*. W przeprowadzonych symulacjach założono, że granica ta wynosi 10 tys. km, co zapewniło, że droga do celu zawsze była znajdowana poprawnie (oznacza to, że w większości wypadków dwa ostatnie łącza satelitarne są wyznaczone algorytmem *min-delay*).

4.4. K-shortest oraz K-shortest z limitacją liczby węzłów

K-shortest to ogólnie znana modyfikacja algorytmu *min-delay*, polegająca na tym, że ruch jest kierowany jedną z k najkrótszych tras. Każda trasa wybierana jest z tym samym prawdopodobieństwem. Dla $k = 1$ działa to oczywiście tak samo, jak algorytm *min-delay*. Badane były jeszcze wartości k równe 2, 3 oraz 4 (oznaczane na wykresach $k2$, $k3$, $k4$).

Łatwo można zauważyć, że algorytm *k-shortest* bardzo często wybiera trasy dużo dłuższe od najkrótszej, wyznaczonej przez oryginalny algorytm *min-delay*. Co więcej, zawierają one często o wiele więcej węzłów pośrednich niż trasa najkrótsza. Stąd wzięło się ogólnie znane ulepszenie, które wprowadza dodatkowe ograniczenie, że wybrana trasa nie może mieć liczby węzłów większej niż trasa najkrótsza. W badaniach symulacyjnych brane były pod uwagę wartości k równe 2, 3 oraz 4 ($k = 1$ oznacza algorytm *min-delay*), oznaczane później na wykresach jako $k2hl$, $k3hl$ i $k4hl$.

4.5. Alr-src

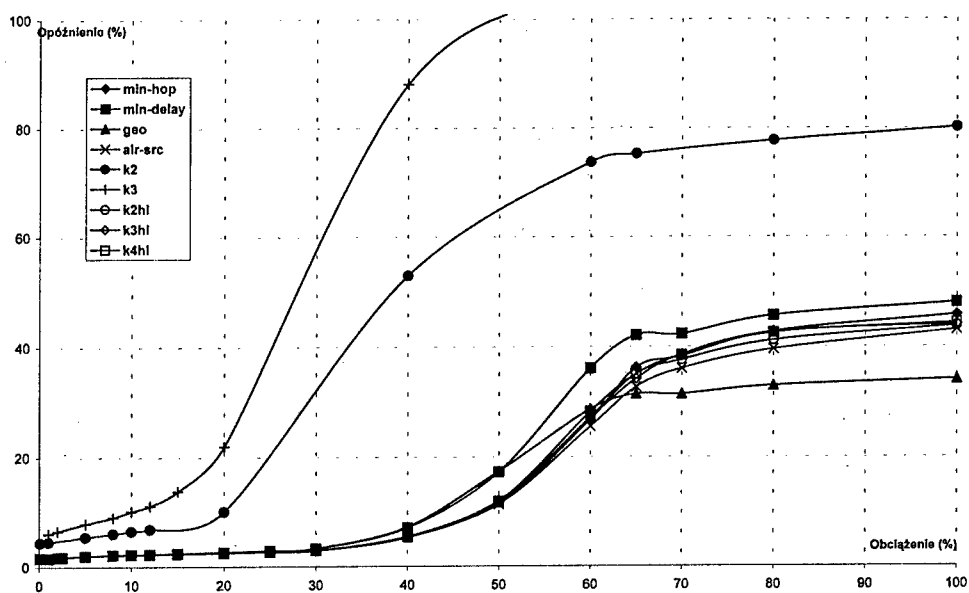
W algorytmie *alr-src* (ang. *Alternative Link Routing with deflection in the SouRCe node*) satelita odróżnia pakiety, które otrzymał od innych satelitów (i musi je przekazać dalej) od pakietów, które odebrał od terminali naziemnych (również musi je przekazać dalej). Te pierwsze pakiety kierowane są trasą najkrótszą pod względem opóźnienia, tak jak w algorytmie *min-delay*, z kolei te drugie są kierowane trasą drugą pod względem opóźnienia, pod warunkiem, że nie ma ona większej liczby węzłów pośrednich – wtedy kierowane są trasą najkrótszą. Algorytm ten został zaproponowany w [4].

5. WYNIKI BADAŃ

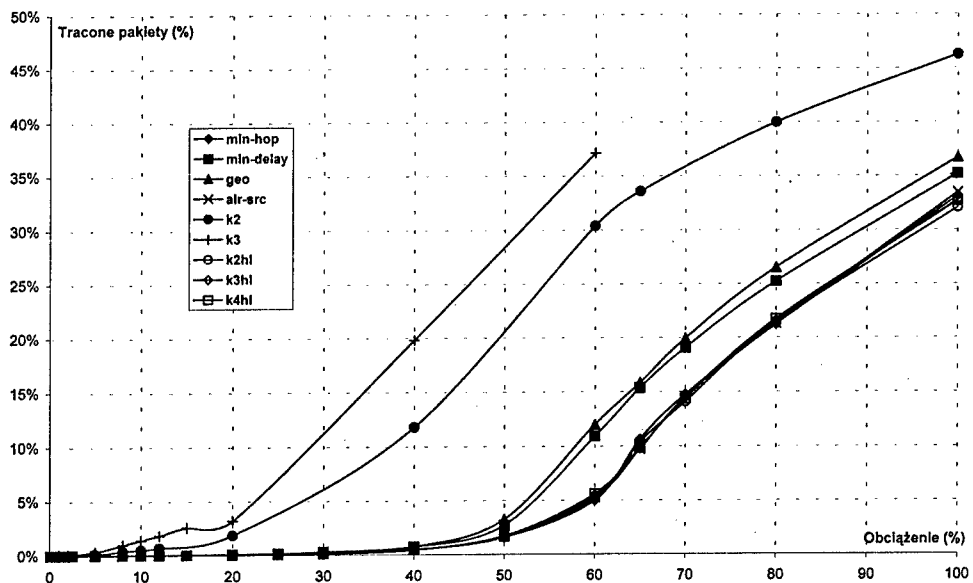
Badania wymienionych algorytmów polegały na zwiększaniu ilości generowanego ruchu od bardzo małej wartości, gdy nie występowały żadne starty pakietów, do coraz większej tak, by coraz bardziej obciążać sieć, co prowadziło do strat pakietów wskutek przepełniania kolejek – nawet do 60% pakietów było traconych. Warto dodać, że wszystkie straty pakietów były spowodowane przepełnieniem kolejek – nie było innych przyczyn ich utraty. Przy zwiększającym się w ten sposób obciążeniu badano:

- średnie opóźnienie doświadczane przez pakiety przesyłane przez sieć;
- liczbę utraconych pakietów.

Wyniki tych badań zobrazowane są na wykresach; rys. 2 przedstawia wykres średniego opóźnienia przy różnych obciążeniach sieci, a rys. 3 przedstawia wykres utraconych pakietów przy analogicznych obciążeniach sieci. Zarówno obciążenie sieci jak i opóźnienia są tu przedstawione w postaci znormalizowanej (maksimum oznaczono przez 100%), gdyż istotne są głównie wartości względne, a wyniki bezwzględne – ze względu na zmiany opisane w rozdziale 2 – nie odpowiadają rzeczywistym wartościom.



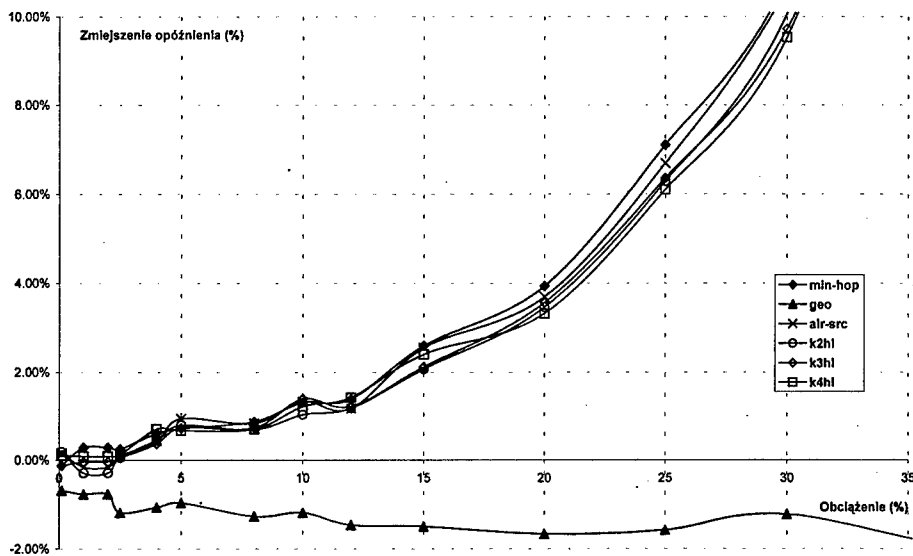
Rys.2. Średnie opóźnienie przy różnym obciążeniu.



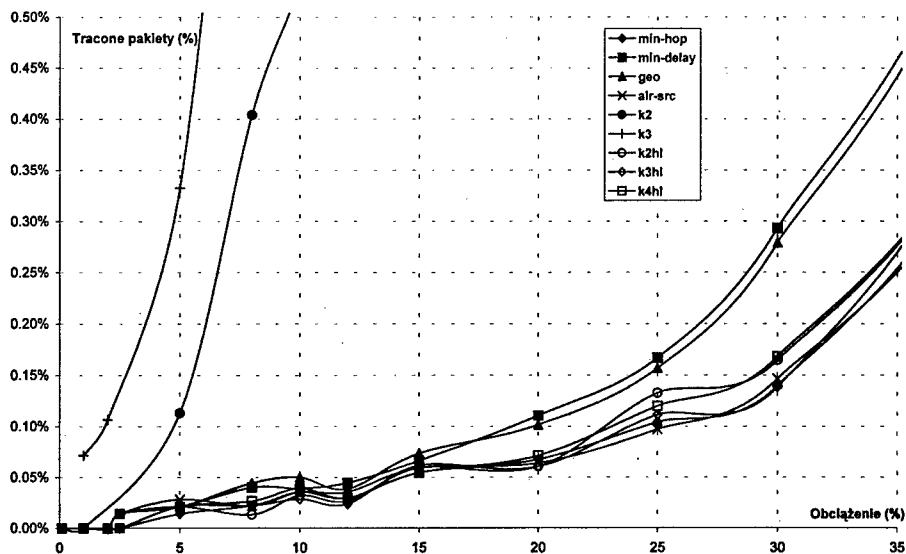
Rys.3. Procentowa utrata pakietów przy różnym obciążeniu.

Zgodnie z oczekiwaniami, opóźnienia stopniowo rosną przy zwiększającym się obciążeniu. W pewnym momencie, gdy obciążenie sięga ok. 60% przyjętej skali, średnie opóźnienie przestaje się prawie powiększać – jest to skutek tego, że kolejki są już całkiem zapełnione. Zwiększa się wtedy prawie liniowo liczba pakietów utraconych.

Jednakże normalnie sieć nigdy nie pracuje w warunkach, w których występują tak znaczne straty pakietów. Z tego względu na kolejnych wykresach skupiono się na obciążeniach, przy których straty utrzymują się poniżej 5%, tj. przy obciążeniu poniżej 35% skali.



Rys.4. Zmniejszenie średniego opóźnienia (zysk procentowy) względem algorytmu *min-delay*.



Rys.5. Procentowa utrata pakietów przy różnym obciążeniu.

Rys. 4 przedstawia zysk jaki dają inne algorytmy pod względem średniego opóźnienia w porównaniu z *min-delay*. Algorytm *geo* jest w większości przypadków gorszy, zaś pozostałe charakteryzują się zwykle mniejszymi opóźnieniami, nawet o 10%. Rys. 5 jest tylko powiększeniem rys. 3 – aby wykresy w analizowanym zakresie były czytelniejsze.

6. WNIOSKI

Wyniki pokazują, iż z badanych algorytmów najgorsze opóźnienia oraz straty pakietów przy dużym obciążeniu występują dla algorytmów z grupy *k-shortest* (dla $k > 1$) bez limitu liczby węzłów – do tego stopnia, że niektórych wyników nie umieszczono na wykresach. Wynika to z tego, że algorytmy te z określonym prawdopodobieństwem wybierają trasy dłuższe od najkrótszej, choć często tak wybrana trasa prowadzi przez dużo więcej węzłów niż najkrótsza. Co prawda obciążenie sieci z pojedynczego połączenia rozkłada się wtedy na więcej węzłów, ale zwiększa się sumaryczne obciążenie sieci i pogarsza się średnie opóźnienie.

Nieco lepsze wyniki uzyskały algorytmy *min-delay* oraz *geo* – kierują one pakiety trasą najkrótszą (*min-delay*) lub wybieraną heurystycznie bliską najkrótszej (*geo*). Jednak kierują one pakiety zawsze jedną trasą, co przy dużym obciążeniu powoduje szybkie napełnianie się kolejek w węzłach leżących na tej trasie, a więc wzrost opóźnień oraz liczby traconych pakietów.

Zdecydowanie lepiej działają algorytmy, które kierują ruch wieloma trasami uwzględniając przy tym liczbę węzłów: *k-shortest* z limitacją liczby węzłów, *min-hop* oraz *alr-src*. Rozkładają one obciążenie na więcej węzłów, przez co zajętość kolejek jest mniejsza, a to z kolei skutkuje mniejszymi opóźnieniami oraz mniejszą liczbą traconych pakietów – np. w zakresie, gdy straty są rzędu 1% pakietów, opóźnienia są o około 6% lepsze niż przy kierowaniu ruchu jedną trasą. Z kolei dla bardzo małych obciążeń sieci algorytmy te uzyskują porównywalne opóźnienia w stosunku do *min-delay*. Z tego powodu można wstępnie ocenić, że mogą one być użyte w sieciach LEO, zarówno przy małym jak i dużym obciążeniu. Dalszych badań wymagają jednak inne aspekty działania tych algorytmów, jak np. kolejność dostarczania pakietów, czy możliwość wykorzystania informacji o aktualnym stanie obciążenia łączy.

BIBLIOGRAFIA

- [1] Hong Seong Chang, Byoung Wan Kim, Chang Gun Lee, Sang Lyul Min, Yanghee Choi, Hyun Suk Yang, Doug Nyun Kim, Chong Sang Kim: *FSA-Based Link Assignment and Routing in Low-Earth Orbit Satellite Networks*. IEEE Transactions of Vehicular Technology, Vol. 47, pp. 1037-1048, Aug 1998.
- [2] Jerome Galtier: *Routing Issues for LEO satellite constellations*. SSGRR 2000 Computer & eBusiness International Conference, L'Aquila, Rome, Italy, July 31 – August 6, 2000.
- [3] Thomas R. Henderson, Randy H. Katz: *On Distributed, Geographic-Based Packet Routing for LEO Satellite Networks*. Proceedings of GLOBECOM, vol. 2, Dec. 2000, pp. 1119-1123.
- [4] M. Mohorcic, A. Svigelj, G. Kandus, Y. F. Hu, R. E. Sheriff: *Demographically weighted traffic flow models for adaptive routing in packet-switched non-geostationary satellite meshed networks*. Computer Networks 43 (2003), pp. 113-131.

EVALUATION OF ROUTING ALGORITHMS IN LEO SATELLITE NETWORKS

Summary

This paper presents results of simulation experiments evaluating several routing algorithms in LEO satellite networks. The experiments take into account non-uniform traffic requirements across the globe. Characteristics of routing algorithms are evaluated under different network loads. Measured parameters are: load distribution, transmission delays and packet losses due to queue overflows.

Tomasz Klajbor, Józef Woźniak

Katedra Systemów Informacyjnych WETI, Politechnika Gdańska

OCENA EFEKTYWNOŚCI PRACY STANDARDU BLUETOOTH

Streszczenie

W publikacji przedstawiono charakterystykę systemu Bluetooth, standardowego rozwiązania dla systemów łączności osobistej. Zaprezentowano również przykładowe wyniki przeprowadzonych badań symulacyjnych. Dokonano oceny przydatności sieci Bluetooth dla różnych topologii i scenariuszy pracy urządzeń końcowych.

1. WSTĘP

Standard Bluetooth został opracowany z myślą o tworzeniu tzw. sieci osobistych PAN (*Personal Area Networks*), tj. sieci o zasięgu do kilkunastu metrów przeznaczonych do komunikacji urządzeń znajdujących się w bezpośrednim otoczeniu użytkownika.

Sieci PAN tworzą zarówno urządzenia przenoszone przez człowieka (np. telefon komórkowy, komputer przenośny), jak i urządzenia stacjonarne (np. drukarka, punkt dostępu do Internetu), znajdujące się chwilowo w jego otoczeniu.

2. CHARAKTERYSTYKA STANDARDU BLUETOOTH

2.1. Historia standardu

W 1994 roku firma Ericsson Mobile Communications rozpoczęła prace badawcze nad standardem Bluetooth. Do maja 1998 roku do firmy Ericsson dołączyli kolejni producenci zainteresowani rozwojem tej technologii (IBM, Intel, Nokia, Toshiba), tworząc grupę *Bluetooth Special Interest Group* (SIG) [1]. Obecnie grupa ta skupia około 3000 firm.

Po opublikowaniu specyfikacji Bluetooth 1.0 (czerwiec 1999 r.) [2], okazało się, że liczne jej niejednoznaczności i błędy utrudniają współpracę urządzeń pochodzących od różnych producentów. W następstwie tego ukazała się errata do wersji 1.0 specyfikacji, a w lutym 2001 r. opublikowano specyfikację Bluetooth 1.1 [3]. W listopadzie 2003 r. ukazała się kolejna modyfikacja standardu oznaczona jako wersja 1.2 [4].

2.2. Techniki transmisyjne

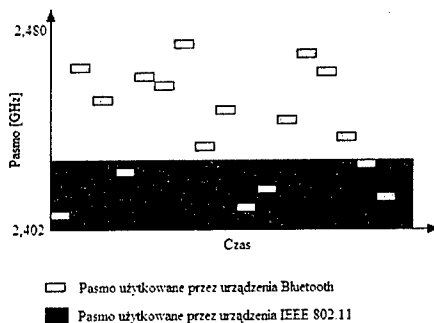
Podczas transmisji w kanale radiowym mogą wystąpić zakłócenia propagacyjne tj. interferencje szerokopasmowe i wąskopasmowe oraz zakłócenia częstotliwościowo-selektywne [5]. W celu minimalizacji wpływu tych zakłóceń, na pracę sieci bezprzewodowych, opracowano transmisji techniki wąsko- i szerokopasmowej.

Rozwiązanie Bluetooth 1.2 [4], podobnie jak wcześniejsze wersje standardu, wykorzystuje nie licencjonowane pasmo ISM (*Industrial, Scientific & Medical Band*) 2.4 GHz. W zależności od kraju zakresy pasma ISM i poziomy emisji radiowej mogą się nieco różnić. Istotnym mankamentem i ograniczeniem ISM jest jednakże współużytkowanie tego pasma przez wiele rodzajów urządzeń.

Aby zapewnić sieciom bezprzewodowym odpowiednią przepustowość z jednoczesną gwarancją pożądanego jakości przekazu, potrzebne są specjalne sposoby transmisji sygnału. Jedną z możliwości oferuje technika szerokopasmowej transmisji FHSS (*Frequency Hopping Spread Spectrum*) [3], stosowana w standardzie Bluetooth 1.1 oraz jej modyfikacja AFH (*Adaptive Frequency Hopping*) [6], zaproponowana w wersji 1.2 standardu. Techniki te polegają na pseudolosowej zmianie częstotliwości podkanałów transmisji danych. Jeśli na pewnej częstotliwości występują zakłócenia lub interferencje uniemożliwiające komunikację, przesyłanie danych jest kontynuowane w następnym skoku, na innej częstotliwości.

Metoda FHSS wiąże się z transmisją w paśmie szerszym niż pasmo sygnału informacyjnego. Dane są transmitowane w postaci ramek – pakietów, z których każdy jest nadawany (i odbierany) na innej częstotliwości (jeden podkanał częstotliwości używany jest w danej szczelinie czasowej o długości T). W przypadku Bluetooth dolna częstotliwość pracy wynosi 2,402 GHz, górna zaś 2,480 GHz. Pasma to podzielono na 79 kanałów (podkanałów) oddalonych od siebie o 1 MHz. Przeskoki między częstotliwościami odbywają się 1600 razy na sekundę. Czas pracy w jednym kanale (tzw. szczelina czasowa albo *Time Slot*) wynosi 625 mikrosekund.

Bezprzewodowe urządzenia wykorzystujące pasmo ISM (np. urządzenia IEEE802.11 i Bluetooth) często mogą wchodzić ze sobą w kolizje – ilustruje to rysunek 1 (zaczepnięty z [6]). W ich efekcie wydajność sieci funkcjonujących na tym samym obszarze wyraźnie spada.



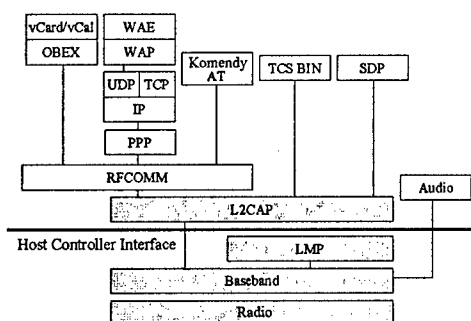
Rys. 1. Rezultat kolizji przy "losowym" FHSS

W celu ograniczenia wpływu wspomnianych interferencji w standardzie Bluetooth 1.2, zaproponowano implementację mechanizmu *Adaptive Frequency Hopping* (AFH). AFH pozwala urządzeniom Bluetooth identyfikować źródła potencjalnych interferencji i wykluczać niedostępne kanały [6].

Właściwy wybór i synchronizacja wykorzystywanych sekwencji kanałów pozwala na pracę w paśmie tylu sieci (teoretycznie) – ile jest wydzielonych kanałów.

2.3. Protokoły w architekturze Bluetooth

Architekturę stosu protokołów standardu Bluetooth [8] ilustruje rysunek 2.



Rys. 2. Warstwowa architektura protokołów Bluetooth

Poniżej przedstawiono krótką charakterystykę protokołów Bluetooth.

Baseband realizuje część funkcji warstwy fizycznej i podwarstwy dostępu do medium warstwy łącza danych (modelu ISO/OSI), definiuje zasady transmisji synchronicznej i asynchronicznej oraz odpowiada za przebieg transmisji (niezawodność komunikacji z wyższymi warstwami).

LMP (*Link Manager Protocol*) służy do komunikacji kontrolnej pomiędzy urządzeniami oraz konfiguruje i zarządza połączeniem warstwy Baseband.

Warstwa L2CA (*Logical Link Control and Adaptation*) korzysta z usług transmisyjnych Baseband. Jest to odpowiednik podwarstwy łącza logicznego warstwy łącza danych modelu ISO/OSI. L2CA odpowiada za multipleksację protokołów warstw wyższych. Odpowiada także za segmentację i składanie danych warstw wyższych oraz usługi połączeniowe i bezpołączeniowe. Warstwa L2CA komunikuje się za pomocą protokołu L2CAP (*Logical Link Control and Adaptation Protocol*).

SDP (*Service Discovery Protocol*) udostępnia informacje o usługach aplikacyjnych obsługiwanych przez urządzenia Bluetooth. Odpowiada za informacje o profilach (i ich usługach) realizowanych przez urządzenia. SDP dostarcza wiedzy o parametrach usług (także i ich adresach) oraz wykorzystywanych protokołach. SDP nie obejmuje procedur realizacji dostępu do usług – procedury te definiowane są indywidualnie dla każdego profilu.

Protokół RFCOMM zapewnia emulację portów szeregowych. RFCOMM jest niezbędnym protokołem podczas komunikacji szeregowej adapterów z urządzeniami zewnętrznymi (nie należącym bezpośrednio do sieci Bluetooth).

Protokół *Telephony Control Binary* (TCS BIN) oraz komendy AT odpowiadają za sterowanie połączeniami telefonicznymi.

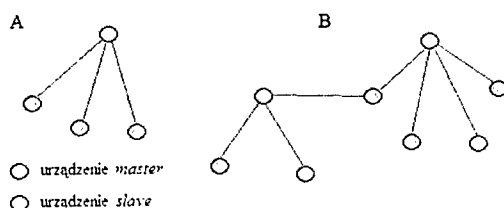
PPP (*IETF Point-to-Point Protocol*) został zaprojektowany w celu transportu pakietów warstwy sieciowej między parą równorzędnych stacji w połączeniach typu punkt-punkt. PPP jest niezbędnym przy transmisji TCP/IP.

Zadaniem protokołu OBEX jest wysyłanie obiektów, takich jak pozycja kalendarza (*vCal*), czy wizytówki (*vCard*) do innych urządzeń. Z kolei

WAP (*Wireless Application Protocol*) umożliwia dostęp do zasobów i usług Internetu przenośnym urządzeniom bezprzewodowym.

2.4. Topologie sieci

Standard Bluetooth przewiduje możliwość tworzenia tzw. „pikosieci” (ang. *piconets*) – grup liczących od dwu do ośmiu urządzeń [4]. „Pikosieć” tworzy jedno urządzenie nadrzędne (*master*) i jedno lub kilka (do 7) urządzeń podrzędnych (*slave*). Urządzenia Bluetooth mogą tworzyć większe struktury, składające się z wielu „pikosieci” – określane jako (z ang.) *scatternet* (patrz rysunek 3).



Rys. 3. Przykłady podsieci Bluetooth (A – „pikosieć”, B – sieć rozproszona – *scatternet*)

Urządzenia w jednej „pikosieci” mogą pełnić rolę nadrzędną jak i podrzędną. Status nadrzędny – *master* – otrzymuje urządzenie, które zainicjowało proces tworzenia sieci. Dopuszczalna jest zmiana statusu urządzeń w sieci (*master-slave switch*).

Przebiegiem transmisji w sieci zarządza urządzenie nadrzędne. Urządzenie podrzędne pragnące skomunikować się z inną stacją, wysyła dane najpierw do urządzenia *master*, które kieruje je do odbiorcy.

2.5. Rodzaje połączeń

Standard Bluetooth 1.2 definiuje następujące rodzaje połączeń logicznych [4]:

- synchroniczne SCO (*Synchronous Connection-Oriented*),
- synchroniczne rozszerzone eSCO (*Extended Synchronous Connection-Oriented*),
- asynchroniczne ACL (*Asynchronous Connection-Oriented*),
- rozsiewcze ASB (*Active Slave Broadcast*),
- rozsiewcze PSB (*Parked Slave Broadcast*).

Pakiety używane w transmisji synchronicznej, SCO, nie zawierają kodu CRC oraz nie są retransmitowane. W przypadku transmisji asynchronicznych, ACL, wykorzystuje się jeden z siedmiu rodzajów pakietów (patrz tablica 1 zaczerpnięta z [7]). Gdy wymagana jest większa szybkość transmisji dopuszcza się tworzenie ramek, które zajmują 3 lub 5 szczelin czasowych. Maksymalna dostępna przepływność (dla ACL) to 723,2 kb/s.

Standard Bluetooth 1.2 definiuje dodatkowo pakiety eSCO [4] (zwane *Extended SCO* lub *Enhanced SCO*), pozwalające na realizację transmisji dźwiękowych wysokiej jakości. Połączenia te wykorzystują różnej długości pakiety EV (z kodowaniem korekcyjnym lub bez). Pakiety EV zawierają sumę kontrolną CRC-16, która w przeciwieństwie do pakietów „połączeń SCO”, pozwala (w ograniczonym stopniu) na retransmisję błędnych jednostek.

Tablica 1

Charakterystyka pakietów stosowanych w transmisji ACL

Typ pakietu	Liczba zajmowanych szczelin czasowych	Liczba bajtów pola danych	Kodowanie korekcyjne	Maksymalna przepływność (symetrycznie) kb/s	Maksymalna przepływność (asymetrycznie) kb/s	
					w dół	w górę
DM1	1	0-17	tak	108,8	108,8	108,8
DH1	1	0-27	nie	172,8	172,8	172,8
DM3	3	0-121	tak	258,1	387,2	54,4
DH3	3	0-183	nie	390,4	585,6	86,4
DM5	5	0-224	tak	286,7	477,8	36,3
DH5	5	0-339	nie	433,9	723,2	57,6
AUX1	1	0-29	nie	185,6	185,6	185,6

Tryb ASB jest używany przez urządzenie nadrzędne do komunikacji z aktywnymi urządzeniami typu *slave*. Z kolei PSB jest stosowany do komunikacji z urządzeniami znajdującymi się w stanie ograniczonego poboru energii.

3. PROFILE UŻYTKOWE BLUETOOTH

Profile definiują sposoby czy też scenariusze wykorzystania technologii Bluetooth. Są one podstawą realizacji tzw. modeli użytkowych. Specyfikacja Bluetooth nie definiuje nam konkretnych obszarów zastosowań tej technologii - to właśnie profile określają zasady realizacji pożądaných funkcji urządzeń.

Profile Bluetooth zostały zdefiniowane w sposób addytywny – implementacja poszczególnych profili może wymagać skorzystania z definicji innych profili. Specyfikacja Bluetooth wymaga od urządzeń nie tylko zgodności z danym profilem, ale także z wszystkimi profilami, od których profil jest zależny.

W tablicy 2 przedstawiono wybrane profile Bluetooth wraz z ich krótką charakterystyką.

Tablica 2

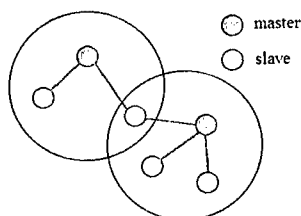
Ważniejsze profile standardu Bluetooth

Nazwa profilu	Charakterystyka profilu
<i>Generic Access Profile</i>	Profil ogólnego dostępu. Odpowiada za procedury wykrywania urządzeń Bluetooth i określa aspekty zarządzania łączem. Jest filarem, na którym opierają się wszystkie pozostałe profile. Specyfikuje także zabezpieczenia kryptograficzne transmisji.
<i>Service Discovery Application Profile</i>	Pozwala na znalezienie informacji odnoszących się do usług udostępnianych przez inne urządzenie Bluetooth. Wykorzystuje protokół SDP.
<i>Serial Port Profile</i>	Odpowiada za utworzenia emulowanego połączenia szeregowego RS-232 z użyciem RFCOMM między parami urządzeń.
<i>Headset Profile</i>	Określa, w jaki sposób bezprzewodowe słuchawki standardu Bluetooth mogą być łączone, aby działały jako interfejs audio dla zdalnego urządzenia.
<i>LAN Access Profile</i>	Specyfikuje zasady dostępu do zasobów sieci LAN lub innej. Umożliwia połączenie dwóch urządzeń poprzez protokoły warstwy sieciowej. Wykorzystuje protokoły PPP i <i>Serial Port Profile</i> .
<i>File Transfer Profile</i>	Umożliwia wymianę plików oraz zdalne zarządzanie katalogami i plikami.
<i>Personal Area Networking Profile</i>	Umożliwia transmisję ramek Ethernet. Wykorzystuje protokół <i>Bluetooth Network Encapsulation Protocol</i> (BNEP), w celu realizacji funkcji mostu do sieci LAN lub realizacji sieci <i>ad-hoc</i> .

Profile Bluetooth zostały stworzone przede wszystkim w celu łatwej implementacji określonego zastosowania technologii – zwanego modelem użytkowym. I tak np. w celu realizacji połączeń wykorzystujących słuchawkę bezprzewodową Bluetooth należy skorzystać z profili: *Headset Profile* i *Serial Port Profile*.

4. BADANIA EFEKTYWNOŚCI PRACY SIECI BLUETOOTH 1.1

Przedmiotem badań była sieć Bluetooth o topologii przedstawionej na rysunku 4. Sieć tę – składającą się z dwóch pikosieci – przebadano pod kątem jej przydatności do obsługi aplikacji typu VoIP oraz transmisji dużych plików danych z wykorzystaniem protokołu FTP (*File Transfer Protocol*). W odniesieniu do aplikacji VoIP przedmiotem eksperymentów było m.in. określenie maksymalnej liczby działających aplikacji zestawionych pomiędzy urządzeniami.



Rys. 4. Ilustracja topologii symulowanej sieci Bluetooth

Eksperymenty przeprowadzono w środowisku symulacyjnym Network Simulator 2 [9] (wraz z modulem Blueware 1.0 [10]).

4.1. Aplikacja VoIP

Aplikacje typu VoIP wspierają przekaz dźwięku poprzez dostępne rozwiązania sieciowe. Informacje dźwiękowe przesyłane są za pośrednictwem pakietów UDP.

Network Simulator 2 posiada implementację UDP wraz z protokołem RTP (*Real-time Protocol*) – umożliwiając tym samym realizację aplikacji VoIP.

W badaniach określano maksymalną liczbę uruchomionych jednocześnie aplikacji pozwalającą na stabilną pracę systemu. Za niestabilną pracę aplikacji VoIP uznawano przy tym sytuację, w której: średnie opóźnienie transmisji pakietów przekraczało 200 ms; odchylenie standardowe średniego opóźnienia transmisji pakietów przekraczało 50 ms; poziom strat pakietów przekraczał 5%.

W eksperymentach dotyczących pakietów VoIP analizowano liczbę poprawnie działających aplikacji w obecności ruchu UDP. W badaniach używano pakietów o długości 60 bajtów (20 bajtów: strumień 8 kbit/s, kodowane 50 pakietów w ciągu sekundy; 40 bajtów: narzut protokołów RTP/UDP/IP).

We wszystkich badaniach przyjęto, że dane są przenoszone przez pakiety Bluetooth Baseband typu DH, nie zawierające kodowania korekcyjnego.

Na podstawie wyników symulacji określono maksymalną liczbę połączeń głosowych, dla których spełnione są wymagania odnośnie jakości transmisji. Zbiórce dane przedstawiono w tablicy 3. Uzyskane rezultaty należy uznać za interesujące i świadczące o przydatności sieci Bluetooth do realizacji połączeń rozmównych.

Tablica 3

Maksymalna liczba połączeń VoIP (spełniających narzucone wymagania)
w zależności od rodzaju połączenia

	Liczba połączeń VoIP
Transmisja pomiędzy urządzeniami <i>master</i> i <i>slave</i> wewnątrz jednej „pikosieci”	22
Transmisja pomiędzy urządzeniami podrzędnymi wewnątrz „pikosieci”	11
Transmisja pomiędzy urządzeniami nadrzędnymi	11
Transmisja pomiędzy urządzeniami podrzędnymi sąsiadującymi „pikosieci”	9

Dla celów porównawczych należy wspomnieć, iż w [11] badano również możliwości przesyłania pakietów zawierających dane dźwiękowe w rzeczywistej sieci Bluetooth. Do tego celu wykorzystano podobnie jak w powyższych badaniach połączenia typu ACL. Połączenia były zestawiane pomiędzy urządzeniem nadrzędnym i podrzędnym w obrębie jednej „pikosieci” dla pojedynczej aplikacji typu VoIP. Wyniki testów, sieci i sprzętu Bluetooth, przedstawione w [11] można traktować jako porównywalne z otrzymanymi w niniejszej pracy.

4.2. Aplikacja FTP

W badaniach dotyczących aplikacji FTP oceniano czas potrzebny na ściągnięcie pliku o długości 100kB pomiędzy poszczególnymi urządzeniami. Protokoły stosu Bluetooth (poczynając od TCP a kończąc na Baseband) powiększają teoretycznie wielkość przesyłanego pliku o 11%. Uwzględniając ten narzut, przesyłanie wspomnianego pliku pomiędzy urządzeniami *master-slave*, zajmuje teoretycznie ok. 1,86 s (zakładając wykorzystanie ok. 496 pakietów DM5). Dodatkowo na długość transmisji pliku wpływ mają czasy, w którym poszczególne protokoły „wpływają” na postać pakietu.

W przeprowadzonych eksperymentach symulacyjnych transmisja wspomnianego pliku trwała ok. 2,66 s. Z kolei czas transmisji w konfiguracjach *slave-master-slave* i *master-slave-master* był zbliżony do 5 s. Niewielka różnica w czasie transmisji pliku była zauważalna pomiędzy urządzeniami nadrzędnymi. Wynika to z faktu, iż urządzenia te obok transmisji zajmują się także „skanowaniem” obszaru, w celu określenia stanów urządzeń i umożliwienia przyłączania nowych użytkowników do sieci.

5. ZAKOŃCZENIE

W systemie Bluetooth 1.1, warstwa Baseband pojedynczego urządzenia, pozwala na zestawienie do 3 połączeń synchronicznych (SCO) [3] dla transmisji dźwięku z kodowaniem 64 kb/s. Alternatywą dla takiego ograniczenia może być (co potwierdziły eksperymenty symulacyjne) zestawianie większej liczby połączeń dźwiękowych w trybie asynchronicznym – ACL. Realizacje połączeń VoIP, w oparciu o transmisję asynchroniczną, jest rozsądnym rozwiązaniem wobec wykorzystania maksymalnie 3 kanałów SCO. Połączenia eSCO (Bluetooth 1.2) pozwalają na bardziej elastyczny wybór przepływności. Rezygnacja w eSCO ze sztywnej przepływności 64 kb/s (jak w Bluetooth 1.1), zapewnia efektywniejszą organizację pracy kanału transmitującego dane dźwiękowe.

Wielu producentów urządzeń Bluetooth implementuje już wersję 1.2 standardu. Wymagane jest aby urządzenia Bluetooth 1.2 były zgodne ze standardem 1.1. Jednakże na rynku długo jeszcze będą dostępne urządzenia wyposażone tylko w kanały SCO (dla dźwięku) i być może w takich przypadkach organizacja kanałów ACL dla strumieni dźwiękowych mogłaby przedłużyć żywotność standardu Bluetooth 1.1.

Należy oczekiwać, że kolejnym impulsem dla ożywienia technologii Bluetooth będzie uruchomienie standardu w wersji 2.0, który umożliwi pracę z szybkością do 2 Mb/s urządzeniom przenośnym i energooszczędnym oraz z szybkością do 10 Mb/s dla wydajniejszych urządzeń, ale już przy wykorzystaniu pasma 5,7 GHz. Optymiści przewidują pojawienie się standardu Bluetooth wersji 3.0, który będzie pracował w paśmie 2,4 GHz i 5,7 GHz, oferując prędkości transmisji w zakresie od 20 Mb/s do 45 Mb/s.

BIBLIOGRAFIA

- [1] Bhagwat P., *Personal Area Networking over Bluetooth*, prezentacja wprowadzająca na konferencji ACM Mobicom 2000, Boston, 6 sierpnia 2000, <http://www.winlab.rutgers.edu/~pravin/publications/papers/Mobicom-handout.pdf>.
- [2] IEEE, IEEE Wireless Standards Zone, Overview, <http://standards.ieee.org/wireless/overview.html>, strona pobrana 28 stycznia 2004 r.
- [3] Bluetooth SIG, Inc., *Specification of the Bluetooth System, Specification Volume 1: Core Version 1.1*, 22 lutego 2001, <http://www.bluetooth.org>, plik pobrany 3 lutego 2003 r.
- [4] Bluetooth SIG, Inc., *Specification of the Bluetooth System, version : 1.2*, 5 listopada 2003, https://www.bluetooth.org/foundry/adopters/document/Bluetooth_Core_Specification_v1.2, plik pobrany 28 stycznia 2004 r.
- [5] Rutkowski D., *Systemy radiokomunikacyjne z rozpraszaniem widma sygnałów i wykorzystaniem podziału kodowo-częstotliwościowego*, Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne nr 9, 1997.
- [6] Hodgdon C., Ericsson Technology Licensing, *Adaptive Frequency Hopping for Reduce Interference between Bluetooth and Wireless LAN*, maj 2003 r., http://www.ericsson.com/bluetooth/files/whitepaper_on_afh_final.pdf, strona pobrana 28 stycznia 2004 r.
- [7] Chen L., Kapoor R., Lee K., Sanadidi M., Gerla M., *Audio Streaming over Bluetooth: An adaptive ARQ Timeout Approach*, <http://saruman.cs.ucla.edu/gerla/pisa-2003-class-material/readings/Audio%20Streaming%20over%20Bluetooth.pdf>, plik pobrany 28 stycznia 2004 r.
- [8] Bluetooth SIG, Inc., *Bluetooth white paper: Bluetooth Protocol Architecture Version 1.0*, 25 sierpnia 1999, http://www.pday.com.cn/technology/bluetooth_documents/ProtArch.pdf, strona pobrana 28 stycznia 2004 r.
- [9] Network Simulator 2, <http://www.isi.edu/nsnam/ns>.
- [10] Blueware, moduł NS-2, <http://nms.lcs.mit.edu/projects/blueware>.
- [11] Kapoor R., Chen L., Lee Y., Gerla M., *Bluetooth: Carrying Voice over ACL Links*, University of California, Los Angeles, <http://www.cs.ucla.edu/NRL/wireless/uploads/rohit-voice.pdf>, plik pobrany 28 stycznia 2004 r.

EFFICIENCY OF THE BLUETOOTH STANDARD

Summary

The paper presents an overview of the Bluetooth 1.1 (and 1.2) standard. It also describes example simulation. Authors analysed effectiveness of UDP – voice and TCP – data streams over ACL links for example pico- and scatternet topologies.

Krzysztof Malek, Dominik Rutkowski, Maciej Sosnowski

Katedra Systemów i Sieci Radiokomunikacyjnych, Politechnika Gdańska

METODY PRZYSPIESZENIA TRANSMISJI DANYCH W SYSTEMIE UMTS

Streszczenie

Referat stanowi pracę przeglądowo-opisową, w której przedstawiono metody umożliwiające zwiększenie szybkości transmisji danych w systemie UMTS. W szczególności omówiono zagadnienie detekcji łącznej sygnałów wielu użytkowników w stacji bazowej i scharakteryzowano podstawowe metody: detekcję dekorelującą, MMSE, PIC oraz SIC. Rozważone zostały także metody dywersyfikacji nadawania z pętlą otwartą oraz zamkniętą. Opisano mechanizmy szybkiego dostępu pakietowego HSDPA, w tym adaptacyjny dobór wartościowości modulacji przy jednoczesnym adaptacyjnym doborze parametrów kodowania kanałowego AMCS, szybki protokół retransmisyjny HARQ oraz transmisję wieloantenową MIMO, realizowaną przy wykorzystaniu zarówno kodowania przestrzenno-czasowego (STC), jak i multipleksowania przestrzennego (BLAST).

1. WSTĘP

Główną cechą charakterystyczną systemu globalnej radiokomunikacji ruchomej trzeciej generacji UMTS jest szeroko rozumiana konwergencja usług, dla których w podstawowym standardzie systemu UMTS (*Release '99*) dostępna będzie transmisja sygnałów o zasięgu globalnym w trybie komutacji kanałów przynajmniej z szybkością 144 kb/s oraz o zasięgu lokalnym w trybie komutacji pakietów z szybkością nie przekraczającą 2 Mb/s [1]. Stosunkowo duża przepływność strumienia danych, z możliwością realizacji przez użytkownika transmisji równoległej przez kanał radiowy, wymagała zastosowania skomplikowanych metod transmisji i odbioru sygnału. Jedną z metod najwcześniej zastosowanych i opracowanych w celu poprawy jakości i zwiększenia szybkości transmisji, stanowią odbiorniki z detekcją łączną sygnałów wielu użytkowników, eliminujące zakłócenia pochodzące od współużytkowników tego samego pasma kanału. Niepożądane efekty wielodrogowości kanału radiowego i powstających wskutek tego zaników sygnału są ograniczane między innymi poprzez technikę dywersyfikacji nadawania.

W ramach rozwoju systemu skupiono się na implementacji transmisji pakietowej opartej na powszechnie stosowanym protokole IP (ang. *Internet Protocol*). W kolejnych wersjach standardu UMTS (*Release 5*) określono metodę szybkiego dostępu pakietowego HSDPA [2] (ang. *High Speed Downlink Packet Access*), skupiającego różnorodne mechanizmy umożliwiające zwiększenie szybkości transmisji w oparciu o adaptacyjne

metody dostosowania się do stanu kanału. W szczególności dostosowanie wartościowości modulacji wraz z szybkim protokołem retransmisyjnym umożliwia zwiększenie średniej przepływności systemu dostępnej dla użytkownika nawet do 10 Mb/s przy zachowaniu jego dotychczasowej pojemności.

Obecnie trwają prace nad standaryzacją nowych rozwiązań techniki transmisji i odbioru wieloantenowego MIMO (ang. *Multiple Input Multiple Output*) w systemie UMTS. Dzięki zastosowaniu kilku anten po stronie nadawczej i odbiorczej można utworzyć wiele równoległych kanałów transmisji, a odpowiednie przestrzenno-czasowe przetwarzanie sygnału oraz wykorzystanie niezależności zaników w każdym z kanałów utworzonych między parami anten T_x - R_x , umożliwia znaczne zwiększenie przepływności.

2. DETEKCJA ŁĄCZNA I DYWERSYFIKACJA NADAWANIA

Jedną z największych przeszkód w osiągnięciu dużych szybkości transmisji w systemach komórkowych stanowi propagacja wielodrogowa w kanale. System UMTS, zaprojektowany z myślą o udostępnieniu jego użytkownikom szerokiego zakresu usług, został oparty na szerokopasmowym interfejsie radiowym z rozpraszaniem widma WCDMA. Interfejs ten z jednej strony gwarantuje bardzo efektywne wykorzystanie dostępnego pasma częstotliwości w oparciu o standaryzowane rozwiązania, z drugiej natomiast stwarza możliwości opracowywania i wprowadzenia nowych metod nadawania i odbioru, pozwalających na dalsze istotne zwiększanie szybkości transmisji w kanale radiowym.

Do najwcześniej zastosowanych zaawansowanych rozwiązań, opracowanych w celu poprawy jakości transmisji w interfejsie WCDMA i zwiększenia szybkości transmisji osiągalnej w systemie UMTS, należy odbiórnik z detekcją łączną sygnałów wielu użytkowników oraz technika antenowa umożliwiająca dywersyfikację nadawania.

2.1. Detekcją łączną sygnałów wielu użytkowników

W rzeczywistym systemie opartym na interfejsie WCDMA nie jest możliwe stosowanie w pełni ortogonalnych ciągów rozpraszających. Wprowadziłoby to bowiem znaczne ograniczenie liczby jego użytkowników oraz niejednakowy stopień rozproszenia. Stąd w systemie UMTS przewidziano zastosowanie charakteryzujących się niepełną ortogonalnością ciągów pseudoprzypadkowych w postaci zespolonych ciągów Golda (tzw. długie ciągi rozpraszające) oraz zespolonych ciągów S(2) (tzw. krótkie ciągi rozpraszające) [7].

Brak pełnej ortogonalności ciągów pseudoprzypadkowych oraz propagacja wielodrogowa uniemożliwia osiągnięcie niezależności przesyłanych sygnałów poszczególnych użytkowników i powoduje występowanie zakłóceń współkanałowych (MAI – ang. *Multiple Access Interference*). Zakłócenia te pogarszają jakość odbioru co wymusza zmniejszenie szybkości transmisji, a więc wyeliminowanie lub ograniczenie tych zakłóceń jest jednym z istotnych celów. Można to osiągnąć przez detekcję łączną sygnałów wielu użytkowników MUD (ang. *Multiuser Detection*). Detektory MUD, w przeciwieństwie do detektorów konwencjonalnych, nie traktują wpływu sygnałów innych użytkowników jak addytywnego szumu gaussowskiego, lecz wykorzystują znajomość tych sygnałów do redukcji interferencji od nich pochodzących [3].

Wśród metod detekcji łącznej, mogących znaleźć praktyczne zastosowanie, wymienić należy detektory liniowe: dekorelujący oraz minimalizujący błąd średniokwadratowy (MMSE – ang. *Minimum Mean Square Error*), a także detektory z równoległą oraz

sukcesywną kompensacją interferencji (PIC – ang. *Parallel Interference Cancellation*, SIC – ang. *Successive Interference Cancellation*).

Pierwsza grupa detektorów realizuje przekształcenie liniowe sygnałów wyjściowych banku korelatorów i umożliwia podejmowanie decyzji łącznej dotyczącej wartości odebranych sygnałów elementarnych wszystkich użytkowników, na podstawie znaku wyniku tego przekształcenia [3]:

$$\hat{\bar{b}} = \text{sgn}(\mathbf{T} \cdot \bar{y}) \quad (2.1)$$

gdzie: $\hat{\bar{b}}$ – wektor estymowanych wartości sygnałów elementarnych nadanych przez każdego z K użytkowników,

\mathbf{T} – macierz przekształcenia liniowego,

\bar{y} – wektor wartości wyjściowych banku K korelatorów odbiornika.

W przypadku detektora dekorelującego, macierz przekształcenia liniowego jest odwrotnością macierzy korelacji ciągów rozpraszających \mathbf{R} :

$$\mathbf{T} = \mathbf{R}^{-1} \quad (2.2)$$

gdzie: $\mathbf{R} = \mathbf{P}^H \cdot \mathbf{P}$ jest macierzą korelacji ciągów rozpraszających,

$\mathbf{P} = [\bar{p}_1 \ \bar{p}_2 \ \dots \ \bar{p}_K]$ – oznacza macierz, w której każda z kolumn jest wektorem reprezentującym ciąg rozpraszający danego użytkownika.

Detektor MMSE stanowi rozwinięcie metody detekcji dekorelującej, pozwalające na uniknięcie możliwego wzmocnienia szumu w procesie odbioru. W tym przypadku macierz przekształcenia liniowego przyjmuje postać:

$$\mathbf{T} = (\mathbf{G} + \sigma_n^2 \mathbf{I}_K)^{-1} \quad (2.1)$$

gdzie: σ_n^2 jest wariancją szumu,

\mathbf{I}_K oznacza macierz jednostkową o wymiarach $K \times K$,

$\mathbf{G} = \sqrt{\mathbf{E}} \cdot \mathbf{R} \cdot \sqrt{\mathbf{E}}$ reprezentuje macierz nieunormowanych wartości funkcji korelacji ciągów rozpraszających, natomiast

$\mathbf{E} = \text{diag}(E_1, E_2, \dots, E_K)$ jest diagonalną macierzą wartości energii poszczególnych odebranych sygnałów.

Ograniczenie interferencji pochodzących od współużytkowników jest także możliwe za pomocą detektorów PIC i SIC. Detektor PIC ma strukturę kilkustopniową. W każdym ze stopni sygnały wszystkich użytkowników ulegają równoległej estymacji, po czym estymaty poszczególnych sygnałów zostają wykorzystane do redukcji interferencji wprowadzanych do sygnałów pozostałych użytkowników. W ten sposób w kolejnych stopniach detektora wytwarzane są coraz dokładniejsze estymaty sygnału nadanego przez każdego z użytkowników. W detektorze SIC z kolei kompensacja interferencji jest realizowana w sposób sukcesywny. Na wejściu detektora sygnały poszczególnych użytkowników są szeregowane według malejącej mocy, a następnie jest estymowany sygnał najsilniejszy. Jego estymata zostaje odjęta od sygnału wejściowego odbiornika, po czym w następnym stopniu detektora sygnał pozbawiony interferencji pochodzących od sygnału najsilniejszego służy jako podstawa wyznaczania estymaty drugiego pod względem oszacowanej mocy sygnału. Operacja kompensacji powtarzana jest w kolejnych stopniach, aż do uzyskania estymaty sygnału najsłabszego. [3]

Ze względu na złożoność procesu odbioru z zastosowaniem rozwiązań MUD, ich wykorzystanie ogranicza się jedynie do stacji bazowych systemu UMTS.

2.2. Dywersyfikacja nadawania

Wielodrogowa propagacja sygnału w kanale radiowym wywołuje niepożądany efekt jego zaników na wejściu anteny odbiornika. Wobec zaawansowanych technik odbioru, przewidzianych dla łącza *w górę*, takich jak odbiór zbiorczy przestrzenny (ang. *Receiver Antenna Diversity*) czy MUD oraz wymagań dotyczących dużych szybkości transmisji w łączu *w dół*, wynikających z asymetrycznego charakteru natężenia ruchu w zastosowaniach internetowych, niezwykle ważne staje się wprowadzenie rozwiązań eliminujących lub silnie ograniczających niekorzystny wpływ zaników sygnału na jego odbiór w terminalu użytkownika. Grupę takich rozwiązań stanowią metody dywersyfikacji nadawania w stacji bazowej. Polegają one na transmitowaniu kilku kopii sygnału do odbiornika terminala, przy czym konieczne jest zapewnienie niezależności zaników w każdym z kanałów utworzonych pomiędzy poszczególnymi antenami stacji bazowej a anteną terminala. Dzięki temu może być zdecydowanie zmniejszone prawdopodobieństwo, że wszystkie kopie sygnału znajdują się jednocześnie w stanie zaniku. Wymagana odległość między antenami zapewniająca niezależność zaników wynosi kilka długości fali nośnej. [4]

Liczne metody dywersyfikacji nadawania można podzielić na dwie podstawowe grupy. Pierwsza z nich to metody ze sprzężeniem zwrotnym, w których sygnał transmitowany jest z poszczególnych anten z odpowiednimi współczynnikami wagowymi. Współczynniki te są dobierane adaptacyjnie w terminalu użytkownika, po czym przekazywane są do stacji bazowej w kanale sterującym, aby uzyskać możliwie najlepszą jakość odbioru. W systemie UMTS przewiduje się stosowanie transmisji dwuantenowej ze współczynnikami wagowymi realizującymi dopasowanie fazy i amplitudy obu sygnałów bądź tylko fazy sygnałów, lub też jedynie włączającymi tylko jedną z obu anten nadajnika (współczynniki wagowe przyjmują w tym wypadku wartości 0 lub 1) [7].

Druga grupa metod dywersyfikacji nadawania to metody pozbawione sprzężenia zwrotnego, w których sygnał zostaje przetworzony w nadajniku w sposób niezależny od aktualnych warunków panujących w kanale radiowym. Sposób przetwarzania sygnału w nadajniku jest tak dobrany, aby umożliwić jak najefektywniejsze wykorzystanie dywersyfikacji po stronie odbiorczej [6]. W systemie UMTS przewidziano zastosowanie dwóch metod dywersyfikacji nadawania bez sprzężenia zwrotnego: z przełączaniem czasowym (TSTD – ang. *Time Switched Transmit Diversity*) oraz przestrzenno-kodową (SCTD – ang. *Space Code Transmit Diversity*) [7]. Pierwsza z nich polega na transmitowaniu sygnału na raz tylko przez jedną z kilku anten nadawczych, przy czym chwile przełączania pomiędzy antenami są z góry ustalone. Druga metoda polega na jednoczesnym transmitowaniu przez każdą z dwóch anten tego samego sygnału dostarczonego do każdej z nich po przemnożeniu przez inny ciąg ortogonalny.

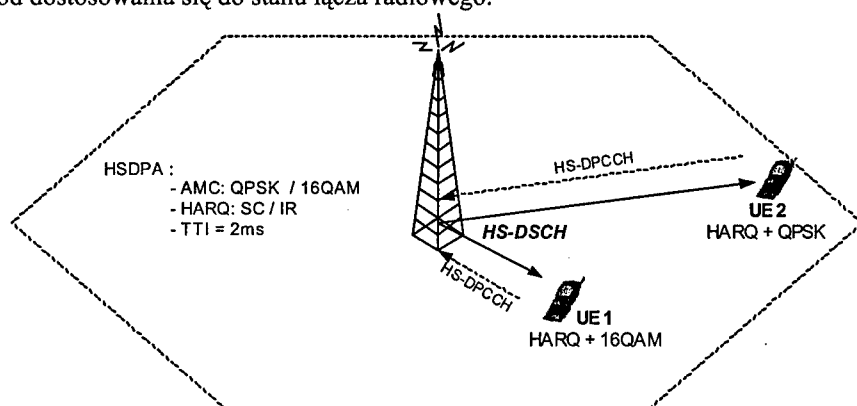
3. EWOLUCJA SYSTEMU UMTS

Wychodząc naprzeciw przewidywanemu rozwojowi aplikacji w systemach radio-komunikacji ruchomej, w ramach procesu ewolucji systemu UMTS określono nowe metody przetwarzania sygnałów głównie pod względem zapewnienia większych szybkości transmisji. Dostęp do dużej przepływności strumienia danych ma zagwarantować

potencjalnemu użytkownikowi możliwość realizacji coraz bardziej wymagających pod względem szybkości transmisji usług multimedialnych.

3.2 Podsystem HSDPA

Nowy podsystem szybkiego dostępu pakietowego w łączu w dół HSDPA (rys.1), ma umożliwić uzyskanie dużych szybkości transmisji poprzez wykorzystanie adaptacyjnych metod dostosowania się do stanu łącza radiowego.



Rys. 1. Symboliczne zobrazowanie pracy podsystemu HSDPA.

W ramach tego podsystemu [2] zdefiniowano nowy typ współdzielonego kanału transportowego w łączu w dół HS-DSCH (ang. *High Speed Downlink Shared Channel*), przy pomocy którego jest realizowana transmisja danych użytkownika w kierunku od stacji bazowej (ang. *Node B*) do terminala ruchomego UE (ang. *User Equipment*). Informacja zwrotna zawierająca dane określające estymacje stanu kanału oraz potwierdzenie lub dyskwalifikację odbioru określonego bloku danych jest dostarczane w kanale zwrotnym HS-DPCCH (ang. *High Speed Dedicated Physical Control Channel*). Do najważniejszych zmian w podsystemie HSDPA w porównaniu do podstawowych funkcjonalności standardu UMTS należy wprowadzenie nowych metod adaptacyjnego dostosowania parametrów transmisji do aktualnego stanu łącza radiowego tj. metod AMC (ang. *Adaptive Modulation and Coding*) i HARQ (ang. *Hybrid Automatic ReQuest*). W ramach metody AMC istnieje możliwość wyboru wartościowości stosowanej modulacji: 2 (QPSK) i 4 (16 QAM) w zależności od położenia stacji ruchomej w obszarze komórki. Uzyskuje się w ten sposób możliwość doboru mocy sygnału w szerokim zakresie i relatywnie większą szybkość transmisji dla modulacji 16 QAM. Metoda HARQ z kolei, czyli połączenie protokołu retransmisyjnego realizowanego w warstwie fizycznej systemu wraz z doбором parametrów kodowania kanałowego, pozwala na szybką retransmisję błędnie zdekodowanych bloków danych. Ponadto opóźnienie retransmisyjne jest zmniejszone poprzez zastosowanie stosunkowo krótkiego czasu między-transmisyjnego TTI (ang. *Transmission Time Interval*) wynoszącego 2ms ($TTI_{min}=10ms$ dla podstawowego standardu). W ogólności można wyróżnić dwa typy protokołu HARQ [2]:

- 1) typ 1 wywodzi się z metody SC (ang. *Soft Combined*) D. Chase'a i polega na kolejnych transmisjach identycznych bloków danych zawierających zarówno ciąg informacyjny

jak i kontrolny. Decyzja o każdym nadanym sygnale elementarnym jest podejmowana na podstawie reguły większościowej obejmującej wszystkie bloki.

- 2) typ 2 – IR (ang. *Incremental Redundancy*) polega na transmisji początkowo głównie danych informacyjnych, a bity nadmiarowe są przesyłane w kolejnych retransmisjach w liczbie pozwalającej na wyraźnie lepsze jakościowo dekodowanie danych. Umożliwia to uzyskanie większej przepływności strumienia danych w korzystnych warunkach propagacyjnych.

Jako kodowanie kanałowe w HSDPA wybrano wysoce skuteczne turbokodowanie o sprawności 1/3, przy czym w ramach jednej z funkcjonalności metody HARQ, polegającej na zmiennym punktowaniu oraz repetycji poszczególnych strumieni bitów, realna sprawność kodowania jest w rzeczywistości zmienna. Przyjęto tu stały współczynnik rozpraszania równy 16 z możliwością transmisji równoległej 15 strumieni danych (ang. *multicode transmission*). Dodatkowo N-kanałowy protokół *Stop and Wait* charakteryzujący się dobieranym odstępem między-transmisyjnym dla poszczególnych użytkowników pozwala uniknąć zbędnej straty czasu wynikającego z procesów retransmisji. Biorąc pod uwagę wszystkie wymienione powyżej czynniki jest możliwa teoretycznie transmisja danych z szybkością przekraczającą 10 Mb/s. Zestawienie możliwych parametrów transmisji i odpowiadające im przepływności zawarto w tabeli 1 [2].

Tabela 1

Zestawienie parametrów transmisji i odpowiadających im przepływności dla podsystemu HSDPA

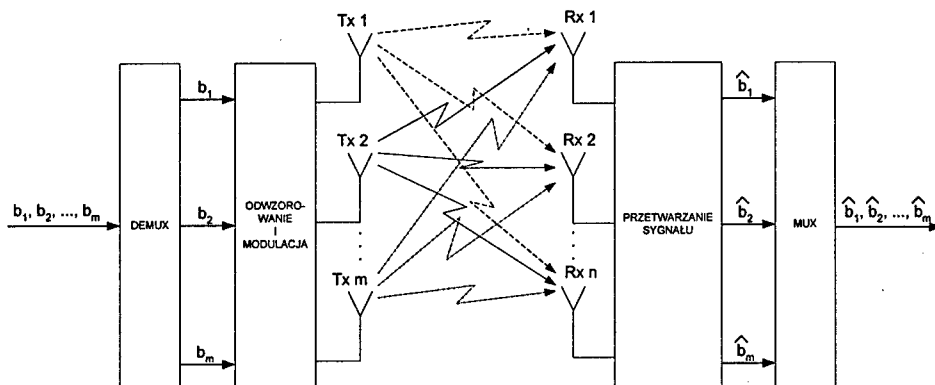
Maksymalna liczba kanałów	Odstęp między-transmisyjny	Liczba bitów w czasie TTI	Typ HARQ	Max szybkość transmisji [Mb/s]
5	3	7300	SC	1,2
5	3	7300	IR	1,2
5	2	7300	SC	1,8
5	2	7300	IR	1,8
5	1	7300	SC	3,6
5	1	7300	IR	3,6
10	1	14600	SC	7,2
10	1	14600	IR	7,2
15	1	20432	SC	10,2
15	1	28776	IR	14,4

3.2 Transmisja wieloantenowa

Terminale użytkowników stają się w miarę upływu czasu coraz bardziej skomplikowanymi urządzeniami dostępu do Internetu. Ograniczenia dotyczące ich gabarytów i złożoności nie są wobec tego już tak ostre, jak w przypadku terminali systemów drugiej generacji, dzięki czemu możliwe staje się zastosowanie podejścia wieloantenowego po obu stronach łącza [5]. Transmisja wieloantenowa (MIMO – ang. *Multiple Input Multiple Output*) to jedno z najnowszych rozwiązań w warstwie fizycznej systemów bezprzewodowych, pozwalające na znaczące zwiększanie jakości transmisji i przepustowości łącza radiowego bez konieczności poszerzania wykorzystywanego pasma częstotliwości. Podejście to polega na stosowaniu kilku anten zarówno po stronie nadawczej jak i odbiorczej, co pozwala na utworzenie wielu równoległych kanałów transmisyjnych. Odpowiednie przestrzenno-czasowe przetwarzanie sygnału oraz wykorzystanie niezależności

zaników w każdym z kanałów utworzonych między parami anten $Tx-Rx$, umożliwia znaczne zwiększenie przepływności. Zdolność odseparowania w odbiorniku sygnałów transmitowanych z poszczególnych anten zależy od niezależności zaników w poszczególnych kanałach. Efektywność transmisji wieloantenowej maleje wobec tego w przypadku bezpośredniej widoczności nadajnika i odbiornika (LOS – ang. *Line of Sight*).

Metody transmisji wieloantenowej można podzielić na dwie grupy: metody multipleksowania przestrzennego (SM – ang. *Space Multiplexing*) zwiększające wydajnie uzyskiwaną w łączu radiowym szybkość transmisji oraz metody kodowania przestrzenno-czasowego (STC – ang. *Space-Time Coding*), zapewniające zysk dywersyfikacji oraz zysk kodowania i wpływające w ten sposób na poprawę jakości transmisji.



Rys. 2. Ilustracja zasady transmisji wieloantenowej z multipleksowaniem przestrzennym

Przykład multipleksowania przestrzennego stanowi metoda V-BLAST (ang. *Vertical – Bell Labs Layered Space-Time Architecture*), której ogólna zasada została przedstawiona na rys. 2. Strumień bitów, pojawiających się na wejściu nadajnika z dużą szybkością, zostaje zdemultipleksowany na m niezależnych podstrumieni, w których szybkość binarna maleje m -krotnie. Wszystkie strumienie są transmitowane jednocześnie przez m anten nadajnika w tym samym zakresie częstotliwości. W odbiorniku, po określeniu macierzy odpowiedzi impulsowych kanałów na podstawie ciągów uczących, sygnały z poszczególnych anten nadajnika zostają odseparowane, po czym jest uzyskiwana estymata nadanego ciągu bitów. Metody odseparowywania poszczególnych sygnałów w odbiorniku są rozwiązaniami analogicznymi do metod detekcji łącznej *MUD* [5].

Wśród metod kodowania przestrzenno-czasowego najważniejsze są metody splotowego kodowania przestrzenno-czasowego STTC (ang. *Space-Time Trellis Codes*), wymagające jednakże stosowania złożonego wielowymiarowego algorytmu dekodowania Viterbiego w odbiorniku oraz metody blokowego kodowania przestrzenno-czasowego STBC (ang. *Space-Time Block Codes*) znacznie łatwiejsze w implementacji ze względu na mniejszą złożoność w procesie odbioru [6].

W chwili obecnej trwają prace nad standaryzacją rozwiązań MIMO w systemie UMTS. Dopracowania wymaga szereg problemów związanych z implementacją transmisji wieloantenowej, takich jak kwestia dopuszczalnej złożoności odbiornika, rozmieszczenia anten czy standaryzacji modelu kanału. Transmisja wieloantenowa została pierwotnie opracowana dla potrzeb wąskopasmowych systemów bezprzewodowych, a więc dla kanału radiowego z płaskimi zanikami. W przypadku kanału selektywnego częstotliwościowo,

z jakim mamy do czynienia w systemach szerokopasmowych, zachodzi konieczność stosowania equalizera kanału wraz z dekodernem przestrzenno-czasowym, co stanowi kolejne wyzwanie dla projektantów [5].

6. ZAKOŃCZENIE

Nieustanny wzrost zapotrzebowania na nowe usługi o coraz większych przepływnościach pociąga za sobą konieczność ciągłego rozwoju standardu UMTS. W chwili obecnej największe oczekiwania związane są z wprowadzeniem podsystemu HSDPA, wspartego rozwiązaniami transmisji wieloantenowej MIMO. Wychodząc naprzeciw rosnącym wymaganiom potencjalnych użytkowników intensywne badania nad nowymi rozwiązaniami dla systemu UMTS będą kontynuowane także po jego uruchomieniu. Badania te będą się koncentrować na nowych metodach coraz bardziej zaawansowanego przetwarzania sygnałów zarówno po stronie nadawczej jak i odbiorczej.

BIBLIOGRAFIA

- [1] Rutkowski D., Sobczak R.: *Usługi w systemie UMTS*, materiały konferencyjne Krajowej Konferencji Radiokomunikacji Radiofonii i Telewizji, Gdańsk 12-14 czerwca 2002, s.21-28.
- [2] Holma H., Toskala A.: *WCDMA for UMTS Radio Access for Third Generation Mobile Communications 2'nd Edition*, John Wiley & Sons LTD, 2002
- [3] Rutkowski D., Sosnowski M.: *Detekcja łączna sygnałów wielu użytkowników w systemie UMTS*, materiały konferencyjne Krajowej Konferencji Radiokomunikacji Radiofonii i Telewizji, Gdańsk 12-14 czerwca 2002, s.315-318.
- [4] Sosnowski M.: *Dywersyfikacja nadawania i jej zastosowanie w systemie UMTS*, materiały konferencyjne Krajowej Konferencji Radiokomunikacji Radiofonii i Telewizji, Wrocław 25-27 czerwca 2003, s.407-410.
- [5] Gesbert D., Shafi M., Shiu D., Smith P.J.: *From Theory to Practice: An Overview of MIMO Space-Time Coded Wireless Systems*. IEEE Journal On Selected Areas In Communications, Vol. 21, No. 3, April 2003.
- [6] Vucetic B., Yuan J.: *Space-Time Coding*. John Wiley & Sons Ltd, 2003.
- [7] 3GPP, Specyfikacja systemu UMTS

METHODS OF DATA RATE IMPROVEMENT IN UMTS

Summary

In this descriptive paper various advanced methods of data rate improvement in UMTS system have been studied. First of all, they concern Multiuser Detection approach in the uplink where decorrelating receiver, MMSE, PIC and SIC solutions have been presented. Both open loop and closed loop Transmit Diversity techniques have also been described as the way of relief of fading effects on the channel. Finally paper discusses the schemes of High Speed Data Packet Access (HSDPA) in UMTS, like Adaptive Modulation and Coding Scheme (AMCS), fast retransmission protocol known as Hybrid Automatic Repeat Request (HARQ) and Multiple Input Multiple Output (MIMO) antenna techniques. Two MIMO approaches have been presented: Space-Time Coding (STC) and Spatial Multiplexing (BLAST).

Andrzej Marczak*, Rafał Niski**

***Katedra Systemów i Sieci Radiokomunikacyjnych, Politechnika Gdańska**

****Instytut Łączności, Gdańsk**

OCENA JAKOŚCI TRANSMISJI DANYCH W INTERFEJSIE RADIOWYM SYSTEMU UMTS Z WYKORZYSTANIEM DOSTĘPNYCH METOD KODOWANIA KANAŁOWEGO

Streszczenie

W referacie została przedstawiona budowa koderów splotowych i turbokodera zastosowanych w systemie radiokomunikacji ruchomej trzeciej generacji UMTS. Przedstawiono i porównano wyniki badań symulacyjnych jakości transmisji danych w interfejsie radiowym z transmisją dwupleksową z podziałem częstotliwości WCDMA/FDD w łączu „w górę” z szybkościami transmisji 144 kb/s i 384 kb/s z wykorzystaniem turbokodera i koderu splotowego o sprawności 1/3. Badania te zostały przeprowadzone w wybranych środowiskach propagacyjnych. Do dekodowania kodu splotowego wykorzystano miękko decyzyjny algorytm Viterbiego, a dekodowanie turbokodu zrealizowano w oparciu o algorytm SOVA.

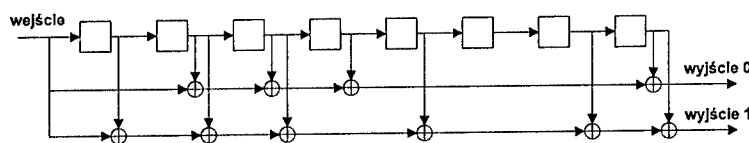
1. WSTĘP

System radiokomunikacji ruchomej trzeciej generacji UMTS (ang. *Universal Mobile Telecommunications System*) będzie zapewniał dostęp do szerokiego zakresu usług związanych z transmisją danych. Będą to usługi o bardzo zróżnicowanych szybkościach transmisji, w tym usługi szerokopasmowe o szybkości dochodzącej do 2 Mb/s. Oprócz dużych szybkości transmisji, wymagana będzie wysoka jakość określona przez prawdopodobieństwo wystąpienia błędów rzędu 10^{-6} - 10^{-8} . Zapewnienie takiej jakości transmisji wiąże się z zastosowaniem odpowiedniego rodzaju kodowania kanałowego w celu zabezpieczenia transmisji przed błędami. W systemie UMTS będą więc stosowane, w zależności od rodzaju kanału transportowego, dwa rodzaje koderów splotowych o sprawnościach 1/2 i 1/3 oraz turbokoder o sprawności 1/3.

2. KODOWANIE SPLITOWE

Przewidziane dla systemu UMTS kodery splotowe zbudowane w oparciu o 9-cio stopniowy rejestr przesuwany przedstawione zostały na rysunkach 1÷2. Dla usług niewymagających wysokiej jakości transmisji stosowany będzie koder o sprawności 1/2, którego wielomiany generacyjne mają następującą postać [1]:

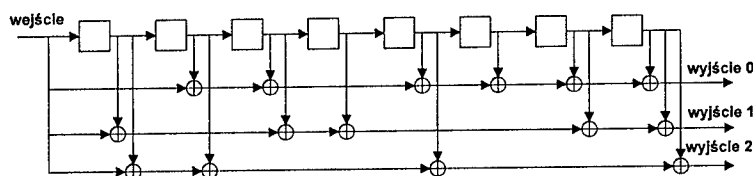
$$\begin{aligned} g_0(u) &= 1 + u^2 + u^3 + u^4 + u^8 \\ g_1(u) &= 1 + u + u^2 + u^3 + u^5 + u^7 + u^8 \end{aligned} \quad (2.1)$$



Rys. 1. Schemat kodera spłotowego (2,1,9)

Dla usług wymagających wyższej jakości transmisji stosowany będzie koder o sprawności $\frac{1}{3}$ i wielomianach generacyjnych [1]:

$$\begin{aligned} g_0(u) &= 1 + u^2 + u^3 + u^5 + u^6 + u^7 + u^8 \\ g_1(u) &= 1 + u + u^3 + u^4 + u^7 + u^8 \\ g_3(u) &= 1 + u + u^2 + u^5 + u^8 \end{aligned} \quad (2.2)$$



Rys. 2. Schemat kodera spłotowego (3,1,9)

Na początku procesu kodowania bloku danych wejściowych, wszystkie rejestry kodera wypełnione są zerami. Ciąg zakodowany powstaje przez pobieranie bitów wyjściowych kolejno z wyjścia 0 i 1, dla kodera (2,1,9), lub z wyjść 0, 1 i 2, dla kodera (3,1,9). Na zakończenie kodowania bloku danych, do kodera wprowadza się 8 bitów końcowych (zerowych), w celu opróżnienia kodera i doprowadzenia go do stanu początkowego.

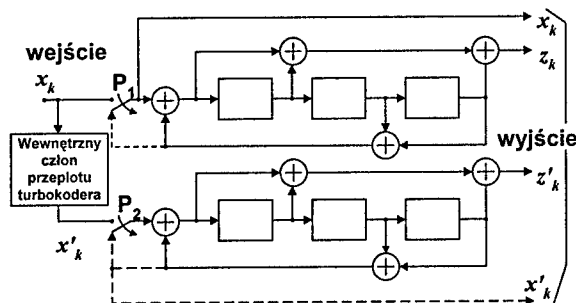
Do dekodowania ciągów kodowych stosuje się algorytm Viterbiego. Jego działanie polega na wyborze odpowiedniej sekwencji bitów, która zapewnia najbardziej prawdopodobną sekwencję stanów (ścieżkę) w grafie kratownicowym kodera.

3. BUDOWA TURBOKODERA DLA SYSTEMU UMTS

Turbokoder zastosowany w systemie UMTS ma sprawność $\frac{1}{3}$. Składa się z dwóch jednakowych, połączonych równolegle, 8-stanowych, rekursywnych, systematycznych koderów spłotowych i członu przepłotu wewnętrznego, przedstawionych na rysunku 3. Funkcję przejścia turbokodera stosowanego w systemie UMTS możemy przedstawić jako:

$$G(u) = \left[1, \frac{1 + u + u^3}{1 + u^2 + u^3} \right] \quad (2.3)$$

Ciąg zakodowany, podobnie jak w przypadku koderów spłotowych, powstaje przez pobieranie bitów wyjściowych kolejno: z wyjścia systematycznego, z wyjścia pierwszego kodera i z wyjścia drugiego kodera [1].



Rys. 3. Schemat turbokodera w systemie UMTS

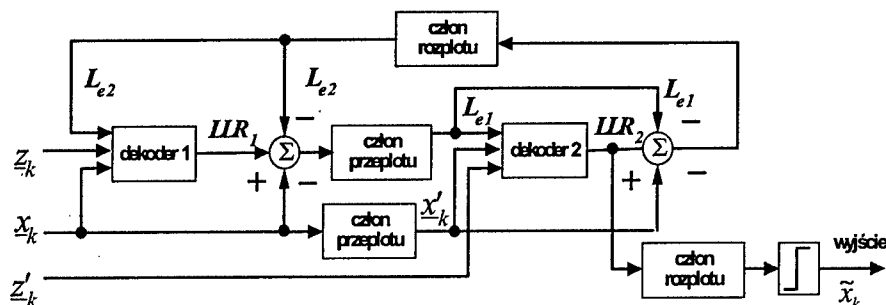
Po zakończeniu kodowania wejściowego ciągu bitów informacyjnych, podobnie jak dla opisanych wcześniej koderów splotowych, następuje proces opróżnienia kodera, czyli doprowadzenie kodera do stanu początkowego. Opróżnienie jest uzyskiwane poprzez wyprowadzenie wszystkich bitów dodatkowych z rejestrów przesuwnych, a na ich miejsce wprowadzenie samych zer. Jest ono realizowane poprzez ustawienie przedstawionych na rysunku 2 przełączników P_1 i P_2 w dolnej pozycji. W wyniku tego na wyjściu turbokodera pojawia się 12 dodatkowych bitów, które są umieszczone na końcu zakodowanego ciągu informacyjnego [1].

W turbokoderze występuje człon przeplotu wewnętrznego, który spełnia bardzo ważną rolę. Oprócz rozproszenia błędów seryjnych, zmniejsza on stopień skorelowania między strumieniem danych wejściowych pierwszego i drugiego kodera splotowego, wchodzących w skład turbokodera [2]. Omawiany przeplot jest przeplotem blokowym, realizowanym w macierzy prostokątnej o wymiarach zależnych od długości ciągu wejściowego. Macierz ta posiada 5, 10 lub 20 wierszy, przy czym maksymalna liczba kolumn wynosi 256. Przeplot składa się z dwóch etapów, przeplotu wewnątrzwierszowego i międzywierszowego [3]. Wynikiem działania członu przeplotu wewnętrznego jest ciąg trafiający na wejście drugiego kodera splotowego turbokodera.

4. BUDOWA DEKODERA TURBOKODU

Budowę dekodera turbokodu przedstawia rysunek 4. Składa się on z dwóch dekode-rów kodu splotowego (dekodera 1 i dekodera 2), członów przeplotu i rozplotu oraz członu ograniczającego na wyjściu dekodera. Członów przeplotu występujące w dekodерze są identyczne z członem przeplotu wewnętrznego turbokodera, natomiast człon rozplotu realizują operację odwrotną.

Dekodowanie turbokodów ma charakter iteracyjny. Polega ono na osobnym dekodowa-niu odebranych ciągów kodowych związanych z obu koderami splotowymi turbokodera [4]. Dekoder 1 dekoduje ciąg odebrany pochodzący z pierwszego kodera z_k , a dekode-r 2 ciąg odebrany pochodzący z drugiego kodera z'_k . Pierwszy dekode-r, w procesie dekodowa-nia, korzysta z tzw. informacji pomocniczej (ang. *extrinsic information*) L_{e2} pochodzącej z drugiego dekodera. Stanowi ona informację o prawdopodobieństwach *a-priori* bitów ciągów informacyjnych [4]. Podobnie drugi dekode-r wykorzystuje informacje L_{e1} pochodzącą z pierwszego dekodera. Dekodery mogą pracować w oparciu o algorytmy bazujące na regule maksymalnego prawdopodobieństwa *a-posteriori* (MAP) [5] lub w oparciu o algorytm Viterbiego z miękkimi wartościami wyjściowymi detektora SOVA (ang. *Soft Output Viterbi Algorithm*).



Rys. 4. Schemat blokowy dekodera turbokodu

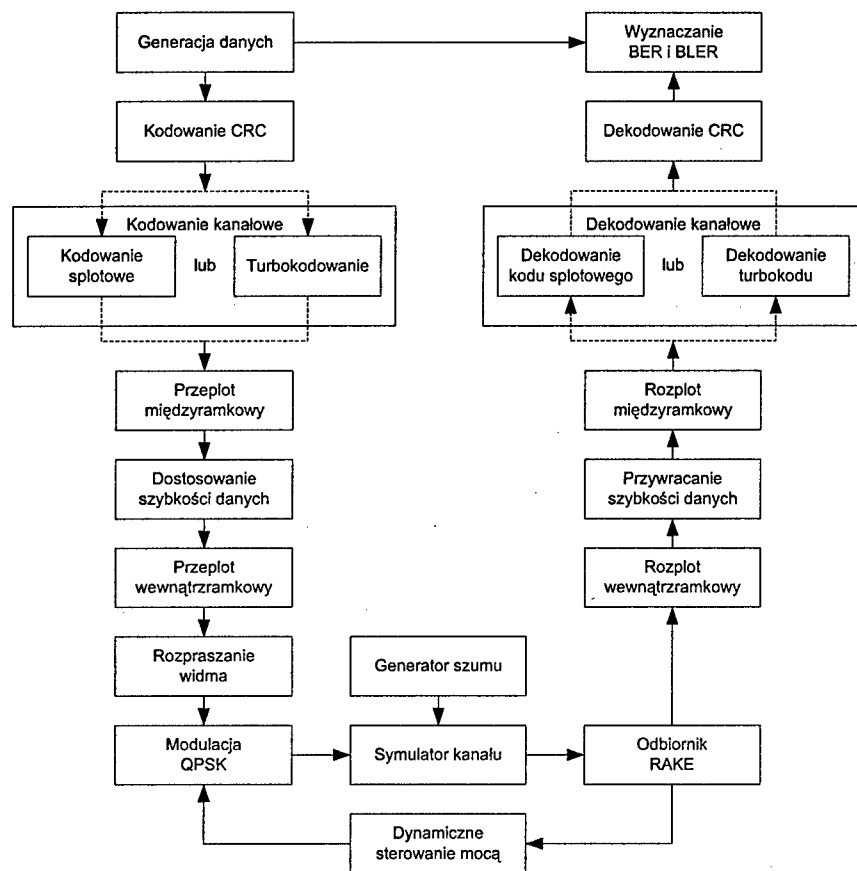
Algorytm SOVA jest zmodyfikowaną wersją algorytmu Viterbiego. Modyfikacje te miały na celu dostosowanie go do wykorzystania w dekodowaniu turbokodów. Po pierwsze zmodyfikowano obliczanie metryk ścieżek w celu uzyskania informacji *a-priori*, kiedy wybrana zostanie ścieżka o maksymalnym prawdopodobieństwie. Drugą modyfikacją jest możliwość dostarczenia miękkiej informacji wyjściowej w postaci logarytmu stosunku prawdopodobieństwa *a-posteriori* LLR (ang. *Log Likelihood Ratio*) dla każdego dekodowanego bitu [6].

5. BUDOWA SYMULATORA

Przedstawiony na rysunku 5 symulator został stworzony w oparciu o własne oprogramowanie symulacyjne. Bloki danych, reprezentujące dane informacyjne, poddawane są w pierwszej kolejności kodowaniu blokowemu detekcyjnemu (CRC), a następnie kodowaniu kanałowemu przy wykorzystaniu kodera spłotowego lub turbokodera. W kolejnych krokach dokonywane są: przeplot międzyramkowy, dostosowanie długości transmitowanego ciągu do szybkości transmisji w kanale oraz przeplot wewnątrzramkowy. Powstały w ten sposób ciąg informacyjny stanowi dane kanału DPDCH, które są poddawane operacji rozpraszania widma. Blok modulatora QPSK odpowiada za utworzenie symboli przenoszących dane przez kanał oraz wzmocnienie nadawanego sygnału powiązane z członem dynamicznego sterowania mocą. W kolejnym kroku sygnał trafia do symulatora kanału.

Wykorzystano tu model kanału radiokomunikacyjnego [7], pracujący w układzie jedna stacja ruchoma i jedna stacja bazowa, umożliwiający symulację zjawisk: propagacji wielodrogowej oraz efektu Dopplera. Na wyjściu symulatora kanału obliczana jest energia sygnału i na tej podstawie generowany jest rozkaz sterowania mocą oraz w zależności od zadanego stosunku E_b/N_0 dodawany jest szum gaussowski, reprezentujący szum termiczny i interferencje od współużytkowników kanału. Po dodaniu szumu sygnał trafia do odbiornika RAKE wyposażonego w układ estymacji przesunięcia fazowego, powstałego wskutek transmisji wielodrogowej. Korekcja fazy odbieranego sygnału dokonywana jest na podstawie nadawanego w kanale sterującym DPCCCH sygnału pilotowego. Następnie w kolejności odwrotnej niż po stronie nadawczej dokonywane są operacje: rozplotu wewnątrzramkowego, odtwarzania rzeczywistej szybkości transmisji danych i rozplotu międzyramkowego. W kolejnym kroku sygnał trafia do bloku dekodera kanałowego. Dekodowanie odbywa się z wykorzystaniem miękkodecyzyjnego algorytmu Viterbiego, w przypadku dekodowania kodu spłotowego oraz algorytmu SOVA, w przypadku dekodowania turbokodu. Następnie dekoderek kodu cyklicznego CRC określa liczbę błędnie odebranych bloków i na tej podstawie wyznacza blokową stopę błędów BLER. W ostatnim członie symulatora

porównuje się zdekodowany sygnał z tym, który został nadany i na tej podstawie określa się elementarną stopę błędów BER.



Rys. 5. Schemat symulatora systemu UMTS/FDD

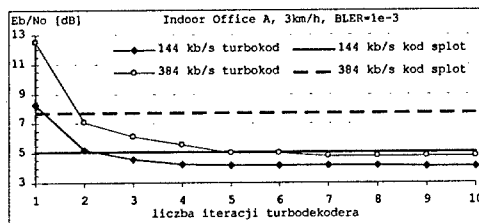
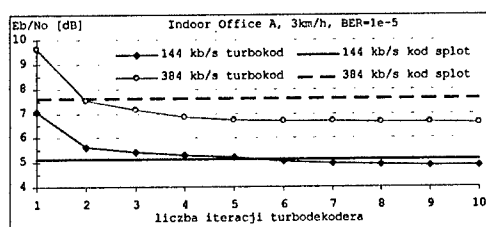
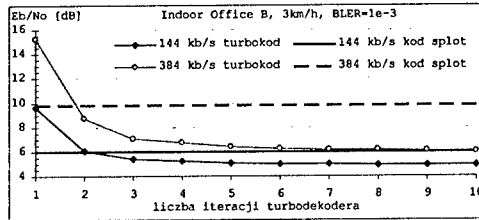
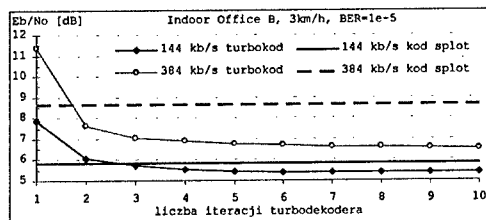
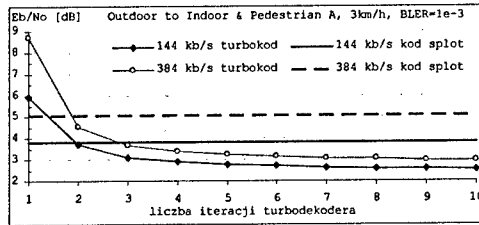
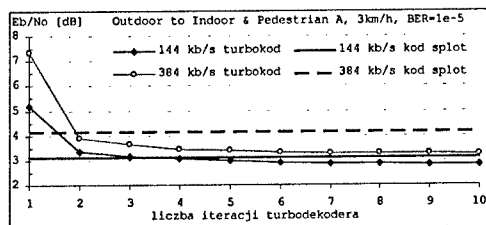
6. PRZEBIEG I WYNIKI SYMULACJI

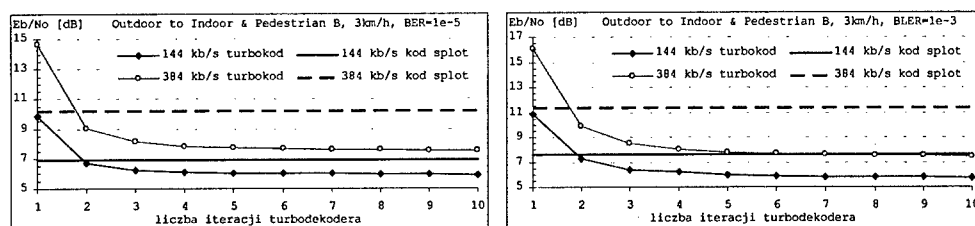
Przeprowadzone badania miały na celu porównanie jakości transmisji danych w interfejsie radiowym z transmisją dwupłową z podziałem częstotliwości (ang. *Frequency Division Duplex*) WCDMA/FDD systemu UMTS w łączu w górę dla kodowania spłotowego oraz turbokodowania. Badania zostały przeprowadzone dla stacji ruchomej przemieszczającej się z prędkością 3 km/h, w środowiskach: wewnątrzbudynkowym (ang. *Indoor Office A i B*) oraz miejskim otoczeniowym (ang. *Outdoor to Indoor & Pedestrian A i B*). W tabeli 6.1 wyszczególniono najważniejsze parametry symulacji. Na rysunkach 6+8 oraz w tabeli 6.2 przedstawiono zależności nominalnych wartości stosunku E_b/N_0 , zapewniających jakość transmisji danych na poziomie $BER=10^{-5}$ i $BLER=10^{-3}$, w funkcji liczby iteracji turbodekodera, dla poszczególnych usług w badanych środowiskach, na tle wartości tego stosunku odpowiadających kodowaniu spłotowemu o sprawności $\frac{1}{3}$.

Tabela 6.1

Wartości parametrów symulacji

Rodzaj usługi	Transmisja danych	
Przepływność	144 kb/s	384 kb/s
Współczynnik rozproszenia widma SF	8	4
Kodowanie detekcyjne	CRC 16	
Kodowanie kanałowe	koder splotowy (3,1,9) / turbokoder	
Dekodowanie kanałowe	miękkodecyczny alg. Viterbiego / alg. SOVA	
Głębokość przeplotu międzyramkowego	40 ms	
Poziom względny wzmocnienia w kanale sterującym (β)	-7,96 dB	-9,54 dB
Środowiska propagacyjne oraz prędkość poruszania się stacji ruchomej	Indoor Office A i B, 3 km/h, Outdoor to Indoor & Pedestrian A i B, 3 km/h,	
Dynamiczne sterowanie mocą	pętla zamknięta ze skokiem 1 dB	

Rys. 6. Zależności E_b/N_0 od liczby iteracji turbodekodera dla jakości usług transmisji danych na poziomie $BER=10^{-5}$ i $BLER=10^{-3}$; środowisko Indoor Office A, prędkość 3 km/h.Rys. 7. Zależności E_b/N_0 od liczby iteracji turbodekodera dla jakości usług transmisji danych na poziomie $BER=10^{-5}$ i $BLER=10^{-3}$; środowisko Indoor Office B, prędkość 3 km/h.Rys. 8. Zależności E_b/N_0 od liczby iteracji turbodekodera dla jakości usług transmisji danych na poziomie $BER=10^{-5}$ i $BLER=10^{-3}$; środowisko Outdoor to Indoor & Pedestrian A, prędkość 3 km/h.



Rys. 9. Zależności E_b/N_0 od liczby iteracji turbodekodera dla jakości usług transmisji danych na poziomie $BER=10^{-5}$ i $BLER=10^{-3}$; środowisko Outdoor to Indoor & Pedestrian B, prędkość 3 km/h.

Tabela 6.2

Nominalne wartości E_b/N_0 [dB] dla poszczególnych usług w badanych środowiskach

Środowisko ¹ propagacyjne	Rodzaj kodowania										
	kod splot	turbokod (liczba iteracji)									
		1	2	3	4	5	6	7	8	9	10
144 kb/s BER=10 ⁻⁵											
Indoor A	5,13	7,05	5,63	5,45	5,31	5,22	5,04	4,95	4,93	4,90	4,87
Indoor B	5,83	7,87	6,06	5,71	5,54	5,43	5,40	5,40	5,40	5,37	5,37
Outdoor A	3,14	5,17	3,37	3,14	3,07	3,00	2,91	2,88	2,87	2,83	2,83
Outdoor B	6,91	9,89	6,68	6,27	6,12	6,05	6,02	6,00	5,95	5,93	5,87
144 kb/s BLER=10 ⁻³											
Indoor A	5,08	8,29	5,23	4,63	4,28	4,15	4,15	4,15	4,15	4,13	4,10
Indoor B	5,97	9,64	6,07	5,43	5,28	5,08	5,00	4,96	4,88	4,88	4,87
Outdoor A	3,80	5,91	3,74	3,11	2,91	2,75	2,71	2,61	2,59	2,56	2,53
Outdoor B	7,59	10,85	7,32	6,42	6,21	5,98	5,91	5,80	5,79	5,79	5,71
384 kb/s BER=10 ⁻⁵											
Indoor A	7,60	9,63	7,54	7,17	6,86	6,74	6,69	6,69	6,65	6,64	6,61
Indoor B	8,63	11,35	7,60	7,03	6,89	6,73	6,71	6,62	6,59	6,55	6,52
Outdoor A	4,17	7,32	3,91	3,65	3,46	3,39	3,34	3,30	3,28	3,27	3,26
Outdoor B	10,23	14,68	9,04	8,15	7,84	7,73	7,67	7,62	7,59	7,56	7,54
384 kb/s BLER=10 ⁻³											
Indoor A	7,70	12,55	7,09	6,12	5,58	5,00	5,00	4,82	4,82	4,82	4,82
Indoor B	9,82	15,22	8,70	7,03	6,72	6,36	6,24	6,14	6,12	6,08	6,00
Outdoor A	5,05	8,72	4,54	3,67	3,40	3,25	3,16	3,07	3,04	2,97	2,94
Outdoor B	11,37	16,10	9,87	8,55	8,05	7,79	7,71	7,59	7,54	7,51	7,47

Zaprezentowane wyniki potwierdzają teoretyczną przewagę jakości turbokodowania nad kodowaniem spłotowym. Przewaga ta jest szczególnie widoczna w środowiskach o gorszych właściwościach propagacyjnych (typu B). Ponadto dzięki zastosowaniu turbokodowania uzyskujemy znaczącą poprawę jakości dla przypadku transmisji danych o przepływności 384 kb/s. Wynika to z faktu większej odporności turbokodu na punktowanie, które w tym przypadku wynosi ok. 17%, co znacznie osłabia działanie kodu spłotowego. Na poprawę jakości transmisji dla tej przepływności znaczący wpływ ma wielkość członu przepłotu wewnętrznego turbokodera, prawie 3-krotnie większa niż dla przepływności 144 kb/s.

Ważnym aspektem turbokodowania jest liczba iteracji w procesie dekodowania. Na podstawie przeprowadzonych symulacji można stwierdzić, iż wystarczająca praktycznie

¹ Oznaczenia: Indoor A/B – Indoor Office A/B, 3 km/h;

Outdoor A/B – Outdoor to Indoor & Pedestrian A/B, 3 km/h.

liczba iteracji powinna zawierać się w granicach od 3 do 6, gdyż poniżej 3 iteracji turbokodowanie jest mniej efektywne od kodowania spłotowego, natomiast powyżej 6 iteracji poprawa jakości związana z kolejnymi iteracjami jest znikoma.

W przeprowadzonych badaniach przyjęto dwa kryteria jakości, które decydowały o nominalnej wartości stosunku E_b/N_0 : $BER=10^{-5}$ oraz $BLER=10^{-3}$. W rzeczywistych warunkach pracy systemu nie jest możliwe określenie elementarnej stopy błędów BER, a jedynie wyznaczenie, dzięki kodowaniu CRC, blokowej stopy błędów BLER. W związku z tym, zastosowanie kryterium BLER jest bardziej właściwe.

Zysk wynikający z zastosowania turbokodu w badanych środowiskach zawiera się w granicach: dla transmisji danych 144 kb/s od 0,9 do 1,7 dB, natomiast dla 384 kb/s od 1,9 do 3,7 dB. Takie różnice wpływają znacząco na zwiększenie pojemności systemu.

7. ZAKOŃCZENIE

Analizując częściowe wyniki badań symulacyjnych, przedstawione w referacie, można zauważyć, że turbokodowanie jest bardziej efektywne od kodowania spłotowego. Uzyskane wyniki pokazały, że zastosowanie turbokodera może być szczególnie przydatne w przypadku usług o dużej przepływności, dla których, w pewnych warunkach, zysk wynikający z zastosowania tego typu kodu sięga prawie 4 dB. Dla takich usług mamy do czynienia z dużym rozmiarem bloku wewnętrznego przeplotu turbokodera, co ma znaczący wpływ na jakość dekodowania.

BIBLIOGRAFIA

- [1] 3GPP TS 25.212 v4.3.0 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; *Multiplexing and channel coding (FDD)* (Release 4).
- [2] Vucetic B., Yuan J. *Turbo codes Principles and Applications*; Kluwer Academic Publishers 2000.
- [3] Marczak A. *Turbokodowanie w systemie UMTS*; Materiały Konferencyjne Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji KKRRiT'2002, Gdańsk 2002.
- [4] Długaszewski Z., Tyczka P. *Twardodecyzyjne dekodowanie turbo-kodów*; Materiały Konferencyjne Krajowej Konferencji Radiokomunikacji, Radiofonii i Telewizji KKRRiT' 2001, Poznań 2001.
- [5] Berrou C., Glavieux A., Thitimajshima P. Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes, Proc. IEEE Conf. On Commun. (ICC'93), Genewa May 1993.
- [6] Woodard J. P., Hanzo L. *Comparative Study of Turbo Decoding Techniques: An Overview* IEEE Trans. On Vehicular Technology vol. 49 No. 6 November 2000.
- [7] Zhang H., *Multipath Fading Channel Model*, Aalborg University, Denmark, 1995.

ANALYSIS OF TRANSMISSION PERFORMANCE OVER UMTS RADIO INTERFACE WITH DIFFERENT CHANNEL CODING METHODS

Summary

In the paper the structures of convolutional – and turbo – codecs used in 3G mobile communication system UMTS have been described. The results of simulations concerning transmission performance using turbo coding and convolutional coding in the uplink of WCDMA/FDD interface have been presented. In the simulations two propagation environments and two data rates 144 kb/s and 384 kb/s have been used.

Paweł Matusz, Józef Woźniak

Katedra Systemów Informacyjnych, Politechnika Gdańska

PERFORMANCE ANALYSIS AND OPTIMIZATION OF THE RADIO LINK CONTROL LAYER IN UMTS

Summary

RLC (Radio Link Control) is one of the sublayers of layer 2 of the radio protocol stack in UMTS (Universal Mobile Telecommunication System). RLC supports three modes of operation: Transparent Mode (TM), Unacknowledged Mode (UM) and Acknowledged Mode (AM). RLC AM performance is highly dependant on the RLC configuration and possibility of dynamic reconfiguration, reflecting changes in the amount and type of handled traffic and radio link quality. Initial configuration of an RLC entity handling a UMTS channel is critical to the overall performance of this channel. The goal of this paper is to present results of RLC performance analysis. The impact of different configuration options on RLC performance has been analyzed and verified by simulation. Based on the obtained results, some optimal configuration guidelines have been proposed. They can be treated as a reference when implementing an effective and optimized UMTS data and control plane.

1. INTRODUCTION

3G (3rd Generation) mobile communication system, referred to as UMTS (Universal Mobile Telecommunication System), is designed to be the successor of 1st generation analog systems and 2nd generation digital systems (such as GSM, Global System for Mobile Communication). In Europe, as well as in Asia and Japan, the main 3G air interface is WCDMA (Wideband Code Division Multiple Access) [1].

All 3G mobile systems, being a part of the IMT-2000 (International Mobile Telephony 2000) standard family, have common design assumptions and goals: mainly to support high user bit rates, advanced QoS (Quality of Service) mechanisms and advanced services generating different types of traffic. Based on these assumptions, standardization organizations have created a set of standards to which all 3G devices should comply. WCDMA for UMTS is being standardized by 3GPP (3rd Generation Partnership Project). A set of documents describes in details all aspects of the UMTS standard including available radio interface modes, services and all protocol stacks.

The radio interface protocol stack is one of the main protocol stacks in UMTS. It handles higher-level data, sent over the radio interface, by appropriately segmenting and reassembling packets, providing low-level ARQ (Automated Repeat Request) support, in-sequence delivery, data compression, flow control, unicast, broadcast and multicast

capabilities and several other functions. The radio protocol stack can be functionally decomposed into three main layers (Fig. 1) [2][3]. Each layer consists of functionally separate protocol entities belonging to different sublayers, and so the second layer can be further decomposed into MAC (Medium Access Control), RLC (Radio Link Control) [4], PDCP (Packet Data Compression Protocol) and BMC (Broadcast Multicast Control) sublayers.

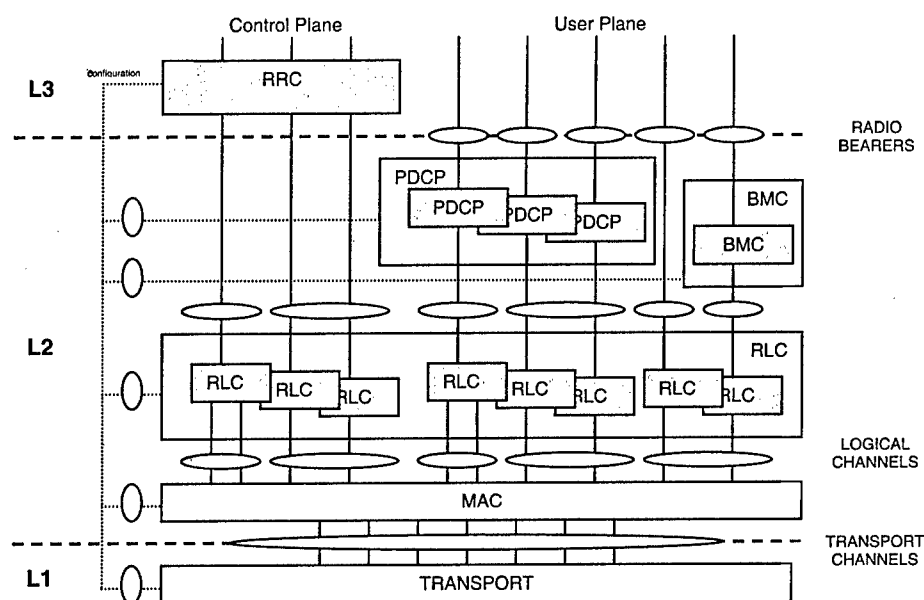


Fig. 1. Functional decomposition of the UMTS radio interface protocol stack

The UMTS RLC layer [4] is an improved version of the RLC layer used in GSM/GPRS and EGPRS [5]. Its main function is to segment RLC SDUs received from the upper layers and create RLC PDUs passed to the lower (MAC) layer, and to assemble RLC PDUs received from MAC to RLC SDUs passed to the upper layers.

RLC supports three modes of operation: Transparent Mode (TM), Unacknowledged Mode (UM) and Acknowledged Mode (AM). In TM, RLC functions are limited to segmentation and reassembly, UM additionally enables detection of missing RLC PDUs and in-sequence delivery. AM is the most sophisticated mode, which additionally implements an ARQ (Automated Repeat Request) mechanism to support retransmissions of lost or erroneous RLC PDUs.

RLC AM performance is highly dependant on the RLC configuration and possibility of dynamic reconfiguration, reflecting changes in the amount and type of handled traffic and radio link quality. Initial configuration of an RLC entity handling an UMTS data channel is critical to the overall performance of this channel. It should take into account average packet length, type of traffic (e.g. constant high or low traffic rate, bursty traffic etc.), asymmetric traffic and possibly other factors to optimally set all RLC parameters. These parameters include RLC PDU size, RLC SDU concatenation, status and poll frequency, status type, status piggybacking and many others. RLC should be flexible enough to adjust most parameters to changes in traffic and link quality over time. Some adjustments should

be done by RRC [6], but some can be done by the RLC entity itself. Correct configuration can maximize the RLC performance, while incorrect configuration can even lead to a situation when RLC will not be operational [7].

The goal of this paper is to present results of RLC AM performance analysis and suggest some guidelines for optimal RLC layer configuration. The results obtained for RLC AM are in general also valid for RLC UM and RLC TM – if applicable, because RLC UM and TM support only subsets of all the functions supported by RLC AM.

2. SIMULATION SETUP

To perform all research and verify expectations, an RLC layer simulator has been used. The simulator fully supports RLC AM, UM and TM and operates in an environment very similar to a real UMTS radio protocol stack.

The overall architecture of the simulated RLC layer is presented in Fig. 2. Appropriate SAPs (Service Access Points) are used to send and receive data from RLC, exactly as it would have been done by the upper and lower layers in a real UMTS system. RLC initial configuration and reconfigurations are done using the control SAP to the simulated RRC (Radio Resource Control) layer.

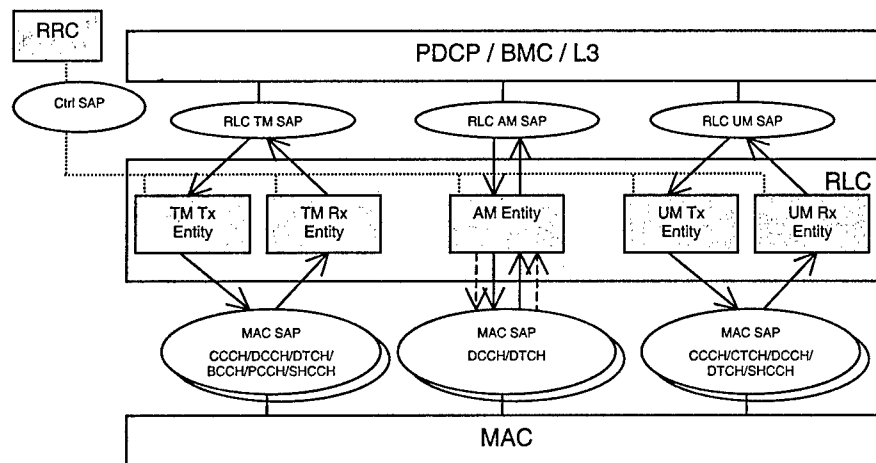


Fig. 2. Overall architecture of the simulated RLC layer

The simulator supports RLC SDU segmentation and reassembly, RLC ARQ mechanism, concatenation, status piggybacking, all control PDU types and maintains a send and reception window. Measurements can be performed in any part of the RLC layer during simulation – header and payload sizes, buffer occupancy, frame timestamps and other statistics are reported and logged. Automated, controlled error injection can be performed.

Two RLC AM entities, located in an RNC (Radio Network Controller) and a connected UE (User Equipment), as shown in Fig. 3, are simulated. The lower layers of the simulator act as MAC and FP (Framing Protocol) layers of the radio protocol stack and the WCDMA (Wideband Code Division Multiple Access) air interface. Some simulations were performed using a single RLC AM entity and some in a scenario presented in Fig. 3, where

one RLC entity was simulated in a UE and one in an RNC and there was a radio bearer established between the UE and the RNC.

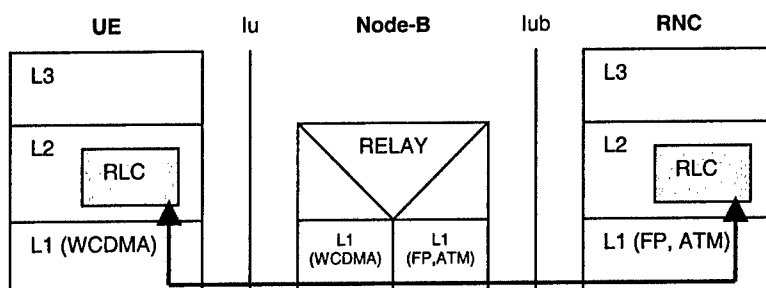


Fig. 3. Connection established between simulated RLC entities

3. SIMULATION RESULTS

The most important feature of RLC AM is the ARQ mechanism. It enables automated, low level retransmissions of lost or erroneous RLC PDUs. The transmitter sends poll requests to the receiver, which is supposed to answer to each poll request with a control PDU carrying a status (STATUS PDU). The receiver may also send unsolicited STATUS PDUs if it is appropriately configured. Each STATUS PDU acknowledges all correctly received PDUs (since the last sent STATUS PDU) and requests retransmission of lost or incorrect PDUs.

Control PDUs consist of a 4-bit header and several SUFIs (Super Fields). SUFIs contain specific control information: they can request entity reset, request moving the reception window, or convey status information. There are 4 types of SUFIs that convey status information: ACK, which acknowledges all correctly received PDUs up to a PDU with a certain sequence number and LIST, BITMAP and RLIST (referred to as NACKs, Negative ACKs), which contain information about sequence numbers of PDUs that need to be retransmitted. A typical STATUS PDU contains zero, one or more NACKs and a single ACK. If all PDUs were received correctly, only the ACK is sent. Otherwise, it is preceded by a NACK (or NACKs) requesting retransmission of PDUs with given sequence numbers.

The LIST, BITMAP and RLIST SUFIs can convey the same information, but the coding of each of them is different (see [4] for details). Depending on the number of PDUs that need to be retransmitted and their distribution among all still unacknowledged PDUs, the optimal (shortest) SUFI should be chosen to reduce the bandwidth consumed by control traffic.

There are two main types of error distributions on a radio link – errors either appear in series (typically as a result of temporary strong deteriorations of SIR) or are sparsely, statistically evenly distributed (typically as a result of constantly low SIR). Fig. 4 presents the length of different STATUS PDU types in case of series of errors of a specific length. Each STATUS PDU consists of a 4-bit header, a variable length NACK SUFI and a 24-bit ACK SUFI. During simulation, such a STATUS PDU was sent every 128 received PDUs and series of errors were injected into a specific number of SUFIs. It can be seen that, on the average, using the RLIST SUFI is the most efficient solution.

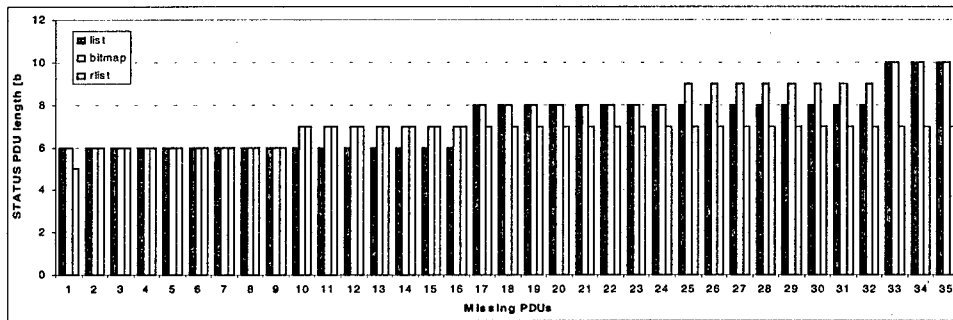


Fig. 4. STATUS PDU length in case of series of errors

The length of different STATUS PDU types in case of statistically evenly distributed errors can be seen in Fig. 5. Errors were injected in such a way, that a certain number of PDUs from the 128-PDU set to acknowledge were incorrect, and the distance (in PDUs) between incorrect PDUs was similar. It can be seen that, on the average, the RLIST SUFI is the shortest. Although for a large number of errors the BITMAP SUFI is shorter (its length is constant, dependant on the distance between the first and last incorrect PDU), a lower error rate is more likely (a very high error rate should force RRC to increase transmission power or reconfigure the radio bearer [7]). Length of the LIST SUFI is a linear function of the number of incorrect, not adjacent PDUs. The RLIST SUFI should preferably be used as the usually shortest SUFI, while usage of the LIST SUFI should be avoided.

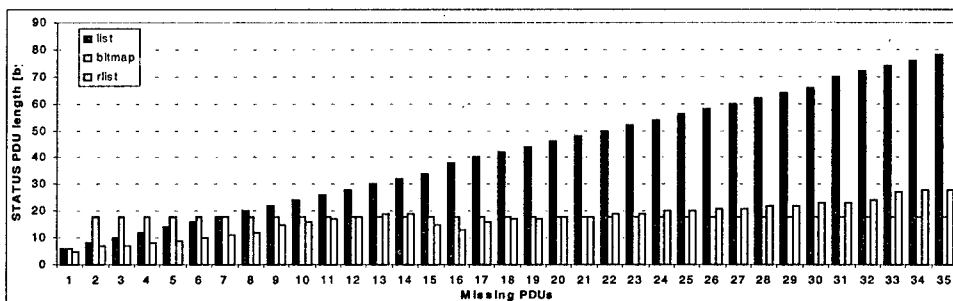


Fig. 5. STATUS PDU length in case of evenly distributed errors

The usage of RLC SDU concatenation is not obligatory and is implementation dependant. If concatenation is used, subsequent RLC SDUs are concatenated in RLC PDUs so, in case of continuous transmission, no space remains unused in RLC PDUs. The usage of concatenation is highly encouraged. If concatenation is not used, each RLC SDU must start from the beginning of a new RLC PDU. Therefore, some of the available bandwidth is wasted by padding inserted after the end of each RLC SDU (unless it completely fills a number of PDUs). This may result in a serious decrease of available user throughput. Fig. 6 depicts the percent of bandwidth consumed by padding in case on not using concatenation, for different RLC SDU lengths and different RLC PDU lengths. As can be seen, not using concatenation is extremely inefficient for short SDUs and long PDUs.

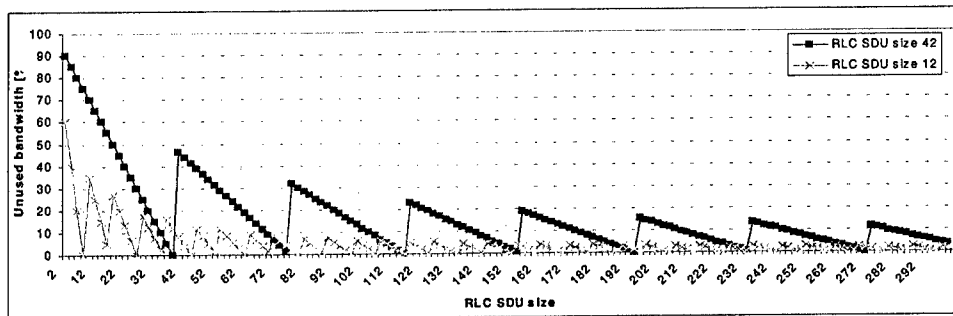


Fig. 6. Bandwidth consumed by padding in case of not using concatenation

In case of using concatenation, the start of each RLC SDU in a PDU is marked by an additional length indicator (LI) in the RLC header. LIs can be 8 or 16-bit long, depending on the length of the PDU (for PDUs longer than 127 bytes 16-bit indicators have to be used). Fig. 7 shows how much of the RLC PDU is used for the RLC header in case of concatenating several RLC SDUs (for 7-bit length indicators) in one PDU. In general, conveying several very short, concatenated RLC SDUs should be avoided, because a serious part of the bandwidth is consumed by RLC headers (the same applies to higher layer headers). Using long PDUs (over 127 bytes) should also be avoided because LIs are twice as long as for shorter PDUs, and in case of errors on the radio link the probability of an error in a shorter PDU is lower than in a longer PDU.

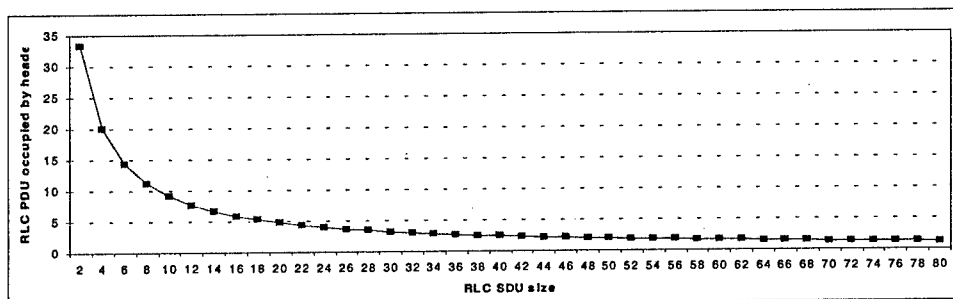


Fig. 7. Length of RLC header for different RLC SDU sizes

The frequency of sending STATUS PDUs largely affects RLC performance. If STATUS PDUs are sent rarely, some bandwidth is preserved, but Rx and Tx RLC buffers must be longer and retransmission time is longer. If STATUS PDUs are sent more frequently, they consume more bandwidth but RLC buffers can be shorter and retransmission time is shorter. The bandwidth consumed by STATUS PDUs in case of different status frequencies (in number of PDUs that are acknowledged by each status) is depicted in Fig. 8.

RLC can support STATUS PDU piggybacking, which is highly recommended. In case of piggybacking, STATUS PDUs can be placed in RLC PDUs instead of padding, after the end of an RLC SDU, although no more SDUs can start after the STATUS PDU in the same RLC PDU. An improvement to the RLC specification would be a possibility to place another SDU after a STATUS PDU inside one RLC PDU. If piggybacking is not configured, each STATUS PDU requires a separate RLC PDU to be transmitted and no

SDUs can be placed in this RLC PDU. As can be seen in Fig. 8, usage of status piggybacking is very efficient, especially if STATUS PDUs are sent frequently.

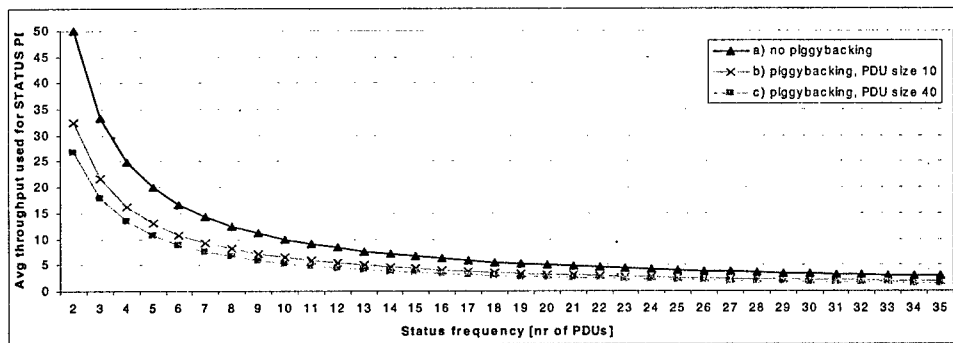


Fig. 8. Average bandwidth consumed by STATUS PDUs:
a) with no piggybacking, b) and c) with piggybacking

Fig. 9 presents the maximum transmission delay, which an erroneous RLC PDU may incur, assuming existence of standard processing delays and no additional delays, such as caused by network congestion or high buffer occupation. Erroneous PDUs received just after sending a STATUS PDU experience such retransmission delays. Delay is presented for 3 simulated RABs (Radio Access Bearers) with different throughputs (identical in uplink and downlink) and different TTIs (Transmission Time Intervals) [8]. The minimum possible delay in the simulated scenario is 2 TTIs and may occur if a status is sent immediately after receiving the incorrect PDU. Processing and transmission delays make it practically impossible to receive the retransmitted PDU after one TTI. It can be seen that, although control traffic overhead is lower for lower status frequencies, the retransmission time (especially for low-bitrate RABs) significantly increases.

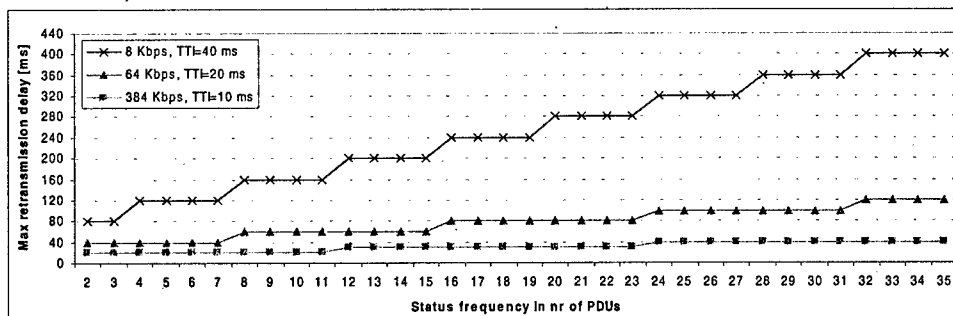


Fig. 9. Maximum retransmission delay

A tradeoff between bandwidth consumed by control traffic and tolerated retransmission delay must be made while configuring and reconfiguring an RLC entity. QoS requirements of conveyed traffic should be considered. Problems with bandwidth consumed by STATUS PDUs can be partially resolved by using different channels for data and control traffic, which is one of the possible RLC AM configurations. In such case,

more radio resources are required on the air interface, but the throughput of the data channel remains constant regardless of the amount of control traffic.

4. CONCLUSIONS

In this paper, the impact of several RLC configuration options and parameters on the overall RLC performance has been presented. Performed simulations allowed verifying some expectations and checking in practice how different parameters affect RLC. Based on the obtained results, either optimal settings were proposed, or configuration guidelines given.

The conclusions are that RLC AM should always use concatenation and status piggybacking, and that the RLST SUFI should be used in STATUS PDUs. RLC PDU size should be small, configuring RLC PDU length more than 127 bytes is inefficient. The frequency of sending statuses should preferably be reconfigurable and should take into account traffic QoS requirements and RAB throughput. Additionally, conveying very short RLC SDUs should be avoided, but this cannot be assured by RLC.

Reconfiguring the status (or status poll) frequency and usage of various available RLC poll and status send modes, including periodical, timer based and RTT estimation modes is a matter of further study.

BIBLIOGRAPHY

- [1] Holma H., Toskala A.: *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. Second Edition, 2002, Wiley & Sons
- [2] Korhonen J.: *Introduction to 3G Mobile Communications*, 2001, Artech House
- [3] 3GPP TS 25.401: *UTRAN Overall Description*, ver. 5.3.0, 6/2002
- [4] 3GPP TS 25.322: *RLC Protocol Specification*, ver. 5.6.0, 10/2003
- [5] 3GPP TS 44.060: *Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control/Medium Access Control (RLC/MAC) protocol*, ver. 5.9.0, 12/2003
- [6] 3GPP TS 25.331: *Radio Resource Control (RRC) Protocol Specification*, ver. 5.7.0, 12/2003
- [7] Zander J., Kim S.L.: *Radio Resource Management for Wireless Networks*, 2001 Artech House
- [8] 3GPP TS 34.108: *Common test environments for User Equipment (UE) conformance testing*, ver. 4.8.0, 10/2003

ANALIZA WYDAJNOŚCIOWA ORAZ OPTIMALIZACJA WARSTWY RLC W SYSTEMIE UMTS

Streszczenie

RLC (Radio Link Control) jest jedną z podwarstw warstwy 2 stosu radiowego systemu UMTS. RLC wspiera trzy podstawowe tryby działania: przezroczysty (Transparent Mode, TM), bezpotwierdzeniowy (Unacknowledged Mode, UM) oraz potwierdzeniowy (Acknowledged Mode, AM). Wydajność RLC AM jest mocno uzależniona od konfiguracji RLC oraz jej zdolności do dynamicznej rekonfiguracji, odzwierciedlającej zmiany w przesyłanym ruchu oraz w jakości łącza radiowego. Celem tej pracy jest przedstawienie wyników analizy wydajnościowej warstwy RLC. Wpływ różnych parametrów konfiguracyjnych na wydajność został przeanalizowany i sprawdzony symulacyjnie. Wyniki tych badań posłużyły do zaproponowania najlepszej, dla różnych rodzajów przesyłanego ruchu, konfiguracji parametrów RLC. Propozycje te powinny być pomocne przy projektowaniu i implementacji efektywnej i wydajnościowo zoptymalizowanej warstwy RLC w systemie UMTS.